ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/319253031

# Investigating Word Affect Features and Fusion of Probabilistic Predictions Incorporating Uncertainty in AVEC 2017

**Conference Paper** · October 2017

DOI: 10.1145/3133944.3133952

CITATIONS
3

READS
263

**10 authors**, including:

Ting Dang
UNSW Sydney
**8** PUBLICATIONS **52** CITATIONS
SEE PROFILE

Zhaocheng Huang
UNSW Sydney
**13** PUBLICATIONS **59** CITATIONS
SEE PROFILE

Sadari Jayawardena
University of Moratuwa
**4** PUBLICATIONS **13** CITATIONS
SEE PROFILE

**Some of the authors of this publication are also working on these related projects:**



Short Duration Language Identification View project



Multimodal affective computing for automated depression analysis View project

All content following this page was uploaded by Brian Stasak on 19 October 2017.

The user has requested enhancement of the downloaded file.

# Investigating Word affect Features and Fusion of Probabilistic Predictions Incorporating Uncertainty in AVEC 2017


Ting Dang
University of New South Wales and Data61, CSIRO, Australia
ting.dang@unsw.edu.au

Brian Stasak
University of New South Wales and Data61, CSIRO, Australia
b.stasak@student.unsw.edu.au

Zhaocheng Huang
University of New South Wales and Data61, CSIRO, Australia
zhaocheng.huang@unsw.edu.au

Sadari Jayawardena
University of New South Wales and Data61, CSIRO, Australia
s.jayawardena@student.unsw.edu.au

Mia Atcheson
University of New South Wales, Australia

Munawar Hayat
University of Canberra, Australia

Phu Le
University of New South Wales, Australia

Vidhyasaharan Sethu
University of New South Wales, Australia

Roland Goecke
University of Canberra, Australia

Julien Epps
University of New South Wales and Data61, CSIRO, Australia



## ABSTRACT

Predicting emotion intensity and severity of depression are both challenging and important problems within the broader field of affective computing. As part of the AVEC 2017, we developed a number of systems to accomplish these tasks. In particular, word affect features, which derive human affect ratings (e.g. arousal and valence) from transcripts, were investigated for predicting depression severity and liking, showing great promise. A simple system based on the word affect features achieved an RMSE of 6.02 on the test set, yielding a relative improvement of 13.6% over the baseline. For the emotion prediction sub-challenge, we investigated multimodal fusion which incorporated a measure of uncertainty associated with each prediction within an Output-Associative fusion framework for arousal and valence prediction, whilst liking prediction systems mainly focused on text-based features. Our best emotion prediction systems provided significant relative improvements over the baseline on the test set of 39.5%, 17.6%, and 29.3% for arousal, valence, and liking. Of particular note is that consistent improvements were observed when incorporating prediction uncertainty across various system configurations for predicting arousal and valence, suggesting the importance of taking into consideration prediction uncertainty for fusion and more broadly the advantages of probabilistic predictions.



KEYWORDS

Depression prediction; dimensional emotion prediction; output-associative fusion; sentiment analysis; uncertainty prediction.


## 1 INTRODUCTION

Two major challenges in affective computing are depression severity prediction and continuous emotion prediction. Whilst the former targets accurate prediction of the severity of depression in patients, the latter aims to predict emotional parameters such as arousal and valence on a per-frame basis, from data such as such as audio, video, and physiological signals.

Clinical depression is among the most prevalent illnesses in the world. Approximately 20% of all individuals will have a bout of depression during their lifetime [1]. Moreover, nearly half of these people will have recurring depression episodes and/or receive inadequate treatment [1, 2]. Currently, many populated nations have a limited number of qualified experts on hand and treatment options to properly address the growing depression epidemic [1, 3, 4].

Continuous emotion prediction that aims to recognize and interpret human affect has seen increasing research interest in the past decades. Predicting the intensity of human affect continuously can provide benefits for interactions between humans and machines, and can additionally serve as clinical aids. For example, a friendly human-computer interface could potentially help individuals who suffer from social anxiety disorder or autism to develop and improve their social skills [5, 6].

*The 2017 Audio-Visual Emotion Challenges and Workshops* (AVEC 2017) offers a standardized multimodal platform for advancing and assessing research into the automated prediction of depression severity and frame-level emotional dimensions in terms of arousal, valence and liking ( i.e. a person expresses positive or negative feelings). A wide range of standard feature sets including audio, video, and text are provided, alongside baseline performance measures for both emotion prediction and depression prediction. This paper presents an investigation into both tasks, with a focus on text-based features and fusion techniques for text-based and acoustic systems in the depression sub-challenge. In the emotion sub-challenge, we focus on high-level linguistic features from phrases phonetic features, and the use of uncertainty in the fusion of sub-system emotion predictions.

## 2 RELATED WORK

## 2.1 Depression

Prior research on the acoustic speech signal has reported many differences between non-depressed and depressed speakers. In particular, studies [7-11] have demonstrated that depressed speakers have flattened prosodic characteristics, decreased vocal intensity, slower speaking rates, and poorer speech intelligibility than healthy speakers. Accompanying the audible speech signal, auxiliary speech behaviors provide additional information regarding a speaker's state of mind and health. For example, auxiliary speech habits include sighs, breaths, laughs, pauses, and repeats. Once again, differences between non-depressed and depressed speakers' communicative auxiliary behaviors have been recorded in studies [7, 12, 13]. Previously, these types of behaviors have proven useful for automatic speech-based emotion and depression classification/prediction [7, 14, 15].

Since spoken language is directly related to quantitative linguistic representations, analysis of high-level language components (e.g. word choice, grammar structure) in the form of Natural Language Processing (NLP) provides insight relating to text-based content [16]. Some common examples of NLP analyses are: word-frequencies within a transcript or compared with other texts, topic coverage range (e.g. word frequency within specific subject categories), *n*-grams (e.g. bigrams, trigrams), and psycholinguistic properties (e.g. age of acquisition, concreteness, imaginability). Investigations utilizing NLP have shown that depression can impact an individual's language content [17, 18].

Conversation sentiment is influenced by context or prior statements, lexical make-up, grammatical complexity, syntactic variables (i.e. adjectives, adverbs, negation, quantifiers), and the type of sentence (e.g. statement, question, exclamation) [19]. Sentiment analysis has gained popularity in text-processing applications (i.e. social media, product reviews, financials) as form of data mining involving automatic analysis of affect, opinions, and attitudes. While sentiment analysis of text-transcripts has demonstrated good performance in predicting affect from natural speech [20], only in the last few years have affect-based ratings also been used for depression severity prediction [21, 22] and classification tasks [23].

## 2.2 Emotion

Extensive research on continuous emotion prediction system has been focused on emotionally informative feature representation and regression modelling techniques. As for the feature representation, a compact set of knowledge driven speech features is gaining in popularity for affective computing, i.e. *The Extended Geneva Minimalistic Acoustic Parameter Set* (eGeMAPS) [24], which is adopted as the baseline features. eGeMAPS features are a few functionals (or statistics) applied on acoustic low-level descriptors (LLDs), which is a common approach to derive emotion-related features. More recently, a Bag-of-Words(BoW) approach, which was proposed originally for text processing, has been applied on LLDs to achieve promising performance for continuous emotion prediction. Moreover, text features for affective computing suggests the effectiveness [25]; the additional Bag-of-Text-Words (BoTW) features achieves promising results for three affective attributes especially for liking.

Recently, Phone Log-Likelihood Ratio (PLLR), which is a set of purely phonetic features, have been successfully applied to continuous emotion prediction, yielding superior performance over the eGeMAPS features with both *Support Vector Regression* (SVR) and *Relevant Vector Machines* (RVM) [26]. This potentially suggests the usefulness of phonetic information, and incorporation of this could aid emotion prediction. However, how to approach the incorporation remains an open question. Motivated by these, a new feature set that combines the acoustic features and phonetic features has been proposed and analyzed.

Another assumption made in the continuous emotion prediction system is that the emotion prediction certainty does not vary over time. However, it has been argued that the emotion may be inherently ambiguous, and it is inappropriate to use hard labels for modelling [27]. Instead, it would be more promising to characterize each emotion prediction as a distribution (commonly Gaussian), where the mean indicates the most probable emotion intensity and the standard deviation quantifies the confidence level relating to that prediction [28]. Most commonly applied regression modelling techniques including SVR [29] and Long Short Term Memory Neural Network [30] optimized the model based on certain objective functions, which are not able to capture the probabilistic distribution of the prediction, while recently proposed approaches including *Gaussian Mixture Regression* (GMR) [31], *Gaussian Process Regression* (GPR) [32], RVM [22, 33] are able to provide distribution-like outputs.

An open question for continuous emotion prediction is multimodal fusion. Adopting multiple regression models potentially reduces the risk of having unreliable predictions. An effective method for multimodal fusion is *Output-Associative RVM* (OA-RVM) [22, 33, 34], which can capture temporal information and dimensional affect dependencies. The predictions from one or more regressors, together with temporal arousal and valence predictions, as well as the input features for further training a regression model can be included within the OA matrices. This motivates us to study the OA-RVM framework as a technique for fusion of multiple modalities and feature types.

# 3 FRONT-END FEATURE EXTRACTION

The AVEC sub-challenges [25] provided several features: Cooperative Voice Analysis Repository for Speech Technologies (COVAREP) [35], eGeMAPS [24], BoW [36], and various video features. Beyond the provided AVEC features, we explored several other additional open-source features described in detail below.

## 3.1 Text-based Features

The *Suite of Linguistic Analysis Tools* (SALAT) [37] was utilized extensively for depression and emotion investigations herein for extracting a range of text-based features. SALAT is an open-source toolkit that contains tools for automatic linguistic and word affect text analysis. Examples of the SALAT tools used

include: *Simple Natural Language Processing Tool* (siNLP) [38] , *Tool for Automatic Analysis of Lexical Sophistication* (TAALES) [16] and *Sentiment Analysis and Cognition Engine* (SEANCE) [39]. This is the first research to use SALAT on spoken transcripts for multi-modal emotion/depression prediction tasks.

*3.1.1 Linguistic*

siNLP [38] was used for linguistic analysis for each transcript. A set of 14-dimensional siNLP features were generated containing information on linguistic attributes such as, the total number of words, unique words, pronouns, articles, negations, and determiners. In addition, TAALES [16] was used to generate a wide range of linguistic lexical proficiency features, such as n-gram index coverage based on variety of texts (i.e. newspaper, fiction, spoken) and psycholinguistic word properties. In total, there were 405 TAALES features included. For details regarding both feature sets, readers may refer to [38], [16].

*3.1.2 Auxiliary Speech Behavior*

Auxiliary speech behavior features consisted of different communication-related actions (e.g. laughs, sighs, breaths, word repeats, average phrase length). For each speaker transcript, the auxiliary speech behaviors were tallied across phrases. The average phrase length was based on the number of words divided by the speaker's total number of phrases. In all, there were 18 auxiliary speech behavior features.

*3.1.3 Word Affect*

In [23], SALAT demonstrated strong depression classification feature performance, which provided text-specific motivation for our AVEC17. SEANCE [39] tool references several popular sentiment indices. However, while all were explored, our experiments herein focus specifically on the ANEW [40], EmoLex [41], SenticNet [42] and Lasswell [43] due to their good preliminary depression prediction performance.

The ANEW features [40] include affective norms for arousal, valence, dominance, and pleasure for ~1k English words. Each of the words, which included nouns, verbs, adjectives, and adverbs, were rated from a 0-10 range; wherein, any rating greater than 5 indicated a positive word token, while lesser were considered negative. Prior research [13,14] has shown that in conversation, individuals with depression exhibit more negative words; hence, why this feature was chosen as a candidate for depression prediction. In total, there were 32 ANEW features, including negation sentence feature consideration (i.e. "*I do not love it*").

The EmoLex features [41] consist of token words that relate to eight specific emotions types (e.g. anger, anticipation, disgust, fear, joy, sadness, surprise, trust). Each emotion type token word list has between roughly 500 to 3k token words. Unlike the ANEW features, the EmoLex feature set comprises of both unigram and bigram token word combinations compiled from multiple sources, such as the Macquarie Thesaurus [44], Google™ n-Gram Corpus [45], and General Index [46]. The EmoLex was of particular interest for depression prediction due to its sadness and joy emotional token word measures. In total, there were 40 EmoLex features.

SenticNet features [42] comprise of nearly 13k token words, which has a sizeable count when compared to the ANEW and EmoLex features. Each token word has evaluated perceptual polarity norms for aptitude, attention, pleasantness, and sensitivity. We choose to evaluate SenticNet for depression because unlike the other word affect feature sets, it implements word multiple degrees of word associations. For example, SenticNet takes into account that the word "*grief*" is often linked to other context-related concepts, such as "*sadness*", *depressed*", "*cry*", and "*death*". Based on these four perceptual norms, SenticNet generates 30 different features.

The Lasswell features [39, 47] are derived from 63 different word lists that are categorized by eight semantic characterizations: affection, enlightenment, power, rectitude, respect, skill, wealth, and well-being. Because of our task of depression prediction, the latter category was of particular evaluation interest. In total, there were 146 Lasswell features.

## 3.2 Phonetically-aware Features

A recently developed approach to incorporate phonetic information into conventional acoustic low-level descriptors (LLDs) is proposed in [48]. Assuming there is a set of frame-level phoneme posterior probabilities $p_t(m), 1 < m < M,$ at frame $t$, and $\sum_{m=1}^{M} p_t(m) = 1$, where $M$ represents the total number of phonemes, any set of $D$-dimensional acoustic features $\boldsymbol{x}_t$ can be made 'phonetically-aware' by weighting them with $p(m)$, giving rise to phoneme-dependent $\boldsymbol{x}_t$ for each phoneme $m$. An exponential parameter of $\alpha \in [0, 1]$ is used to control the amount of phonetic information $p_t(m)$ incorporated: $\alpha = 0$ means no phonetic information is considered, while $\alpha = 1$ employs the full phonetic information. Weighting $\boldsymbol{x}_t$ by $p_t(m)$ can be regarded as the acoustic $\boldsymbol{x}_t$ dependent on the presence of phoneme $m$, which concurrently characterizes linguistic contents as well as LLDs. The final phonetically-aware features are formed by concatenating all the phoneme-dependent acoustic features.

# 4 DEPRESSION SYSTEMS

For our first depression system, individual feature sets were independently evaluated using well-established regression methods, Gaussian Staircase Regression (GSR) and SVR (Section 5). This helped to determine which features and regression methods performed best for the development data. As shown in Fig. 1, we also proposed experimentation using decision-level fusion. This approach specifically used our three best performing features (i.e. *Mel-frequency cepstral coefficients* (MFCCs), *Actual Unites* (AU), ANEW) for audio, video and text.

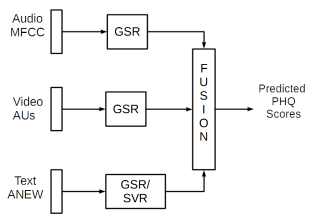


**Figure 1: Proposed systems for depression prediction.**

# 5 EMOTION SYSTEMS

## 5.1 System Overview

The AVEC 2017 emotion sub-challenge provides baseline feature sets of three modalities, i.e. audio, video, and text. The audio features include a set of 88-dimensional eGeMAPS features and a set of BoW representation applied on LLDs. Video features include a set of 121-dimensional features describing facial position and expression of subjects, and a set BoW representation applied on the video features. For text-based features, again a set of BoW representation was applied on transcripts to form frame-level 521-dimensional features. Accompanying the provided features, two novel types of features were applied for emotion prediction, i.e. the word affect features (Section 3.1) and the phonetically-aware acoustic features (Section 3.2).

Based on experiments reported in [25], the emotion dimension liking seems to be strongly associated with text, but barely correlated at all with the audio and video modalities. Therefore, in this paper, systems for liking herein adopted only text-based features, whilst all three modalities were used for arousal and valence. An overview is illustrated in Fig. 2.

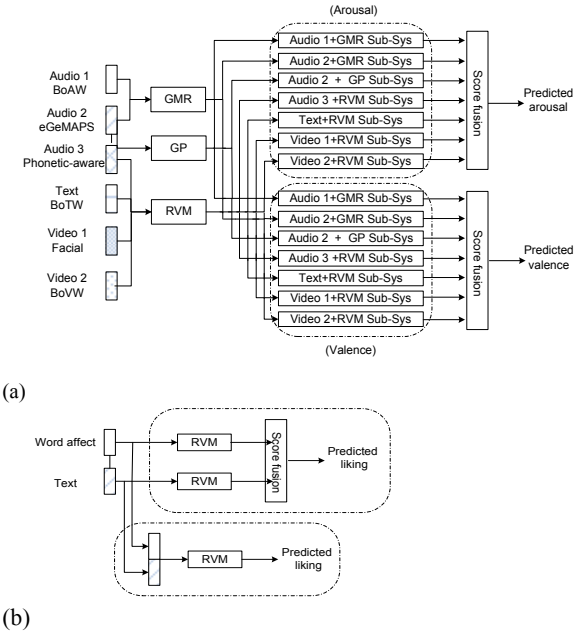


**Figure 2: Overview of emotion prediction systems; (a)Arousal/valence system (b) Liking system**

As shown in Fig. 2(a), the proposed system for arousal and valence comprises 4 acoustic subsystems, 1 text-based subsystem, and 2 video subsystems. For regression modeling, GMR, RVM, and GPR were investigated with empirical selection of front-end features. GMR was found effective for emotion prediction using acoustic features [31], and therefore it was applied to eGeMAPS and bag-of-audio-word (BoAW) features herein. RVM was used on text-based features, phonetic-aware acoustic features and two sets of video features, since it enforces sparsity in features, which can help prevent the overfitting issue of high-dimensional features. GPR was applied to eGeMAPS features due to its ability to provide fully Bayesian modelling of output distributions that simultaneously take into account both temporal and feature-space relationships. Predictions from the seven subsystems were fused via OA-RVM to generate final arousal and valence predictions (Section 6 and 7).

Fig. 2(b) shows an overview for the liking system. In terms of word affect feature extraction, transcripts were translated from German to English using Google Translate™. This approach was taken because the text-based tools in SALAT were designed for English. The conversion is sensible, since there is ~60% lexical similarity between German and English, despite the grammatical differences between these two languages. Linguistic and word affect features were extracted at phrase level using SALAT, which is followed by Shape-preserving piecewise cubic interpolation to align with ground truth ratings, i.e. 0.1 sec per frame. The proposed system concatenated the word affect features and BoTW features to predict liking via RVM or OA-RVM.

## 5.2 Fusion of Probabilistic Predictions

*5.2.1 Quantifying Uncertainty*

GMR and GPR are utilized to quantify the uncertainty owing to their capability of the probabilistic output. These two methods have different, perhaps complementary strengths: GMR aims to capture the joint distribution of features and labels, where the prediction in each frame is represented as a distribution giving the confidence level of the prediction, while GPR can be configured to model the relationship of the current frame and all other frames, which captures the temporal information.

GMR, which aims to model a joint distribution of features and ground truth emotion ratings as a Gaussian Mixture Model (GMM), is an effective regression approach for predicting emotion [8]. To perform prediction, the conditional probability of arousal and valence predictions given features can be estimated and approximated as a single Gaussian distribution by selecting the dominate mixture $\hat{m}$ [28]. Then this single Gaussian distribution corresponding to the dominant mixture is utilized to approach the overall distribution. The expectation and the standard deviation of the Gaussian distribution are regarded as the arousal/valence prediction and uncertainty prediction.

GPR is a Bayesian method which represents distributions over output functions as collections of random variables, which are known as *Gaussian Processes* (GPs). GPs assume any finite subset of the variables have joint Gaussian distribution. For inference purposes, these GPs are defined using a covariance function, which specifies the covariance between two output points (in this case, each point is an emotion dimension value, i.e. arousal and valence at a particular time) in terms of its corresponding predictive input features.

To address this task, we used the temporally-aware additive covariance structure. This model allowed us to specify separate covariance functions to address both the input-output relationships between the features and corresponding emotion ratings, and the temporal interrelationships between emotion dimensions within a given recording. We used the additive kernel presented in [49] to

model the input-output relationships, and a squared-exponential kernel to impose a condition of temporal smoothness on the output distribution.

GMR and GPs are able to output a probabilistic distribution as a Gaussian distribution in both cases. The mean and standard deviation of the Gaussian distribution are treated as an estimation of the emotion attribute prediction and the uncertainty prediction, which were used for multimodal fusion.

*5.2.2 Fusion of probabilistic predictions*

Standard fusion approach only takes into account the emotion attribute predictions obtained from multiple regressors. Instead, our proposed fusion strategy aimed to additionally incorporate the uncertainty predictions, which provided additional information of inherently emotion variability and can be learned by regression techniques.

As shown in Fig 3, emotion attributes predictions from GMR and GP are represented by dash line and solid line respectively. The corresponding posterior distributions of GMR and GP predictions at one specific time step are two Gaussian distributions, with the uncertainty predictions represented by standard deviations $\boldsymbol{\sigma_1}$ and $\boldsymbol{\sigma_2}$.

OA fusion and OA regression were used as our fusion strategy. Standard OA fusion generated the OA matrix by concatenating the original features, arousal predictions and valence predictions represented by $\mathbf{m_1}$ and $\mathbf{m_2}$ of the 1st stage regressors , which is served as the input features of the 2nd stage regression model. In our proposed system, we additionally concatenated the uncertainty predictions of $\boldsymbol{\sigma_1}$ and $\boldsymbol{\sigma_2}$ with the standard OA matrix in a frame basis, which incorporated the uncertainty information in the 2nd stage regressor.

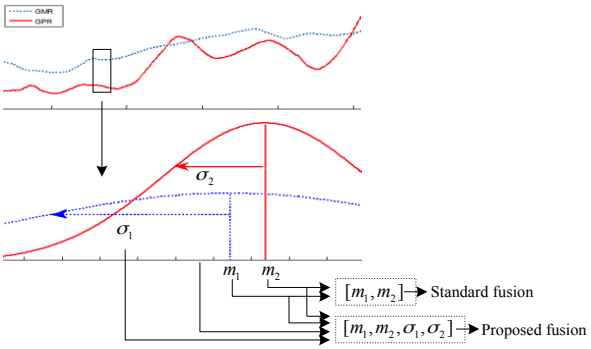


**Figure 3: Fusion of probabilistic predictions: GMR probabilistic prediction (dash), GPR probabilistic prediction (solid); $m_1$ and $m_2$ means mean values and $\sigma_1$ and $\sigma_2$ means standard deviation**

# 6 EXPERIMENT SETTINGS

## 6.1 Database

For the AVEC 2017 depression sub-challenge, the *Distress Analysis Interview Corpus* (DAIC) [50] was used. The DAIC database was created to research spoken language interactive behaviors. It consists of the word-level human transcriptions including auxiliary speech behaviors, and Patient Healthcare Questionnaire (PHQ-8) [51] scores ranging between 0-24. The AVEC 2017 emotion sub-challenge was assessed on the subset of *Sentiment Analysis in the Wild* (SEWA) database [52], a spontaneous audio-video database collected 'in the wild'. The subset used in the emotion sub-challenge contains audio recordings from 64 subjects. Three emotion dimensions, i.e. arousal, valence and liking were rated by 6 annotators every 100 ms. The final ground truth emotion ratings were provided. More details regarding the emotion and depression sub-challenge databases can be found in [50, 52].

## 6.2 Depression System Configuration

For the depression sub-challenge, three modalities were investigated for predicting PHQ scores, i.e. acoustic, video, and text-based features. For the generalization purpose, we combined the training and development data, and used 3-fold cross validation in all experiments

In the acoustic subsystem, the 13-dimensional MFCC, alongside their delta (Δ) and delta-delta (ΔΔ) features were used, 39 dimensions in total. The backend regression approach for the acoustic system is GSR (Section 4) [53]. GSR is an extension of GMM. Instead of creating two separate GMMs for low/high depression classes, GSR uses multiple Gaussians to form the GMM. We used the GSR model described in [53] but without adaptation. The class ranges for the low depression class were 0-5, 0-7, 0-11, 0-16, 0-19 and the complement of these ranges was taken as high depression class. These thresholds were decided empirically. GSR outputs a *Log Mean Likelihood Ratio* (LMLR) for each low and high class pair. Resultant LMLR vectors, which were calculated on a per-file basis, were then used to train a linear SVR. Epsilon parameter of the SVR model was optimized for the range of [10-0.01]. The same GSR framework was also applied to the provided AUs feature set.

For the text-based word affect features, a linear SVR was used as back-end regression model to predict PHQ scores on a per-file basis. The adopted word affect features include 32-dimensional ANEW, 40-dimensional EmoLex, and 146-dimensional Lasswell, all extracted from SALAT SEANCE toolkit on a per-file basis, i.e. one single feature vector per file.

Two standard performance metrics, Root Mean Squared-Error (RMSE) and Mean Absolute Error (MAE) were used to evaluate the overall predictive accuracy.

## 6.3 Emotion System Configuration

All parameters were tuned based on cross validation and in turn used accordingly on the development set and test sets. Delay was tuned within the range of [0,10] sec with an increment of 1 sec. Finally, a 2-sec delay and a 1-sec delay were applied to the audio and video modalities respectively, except that 3-sec delay was used in the GP subsystem. A binomial filter with length of 41 frames was applied for smoothing final predictions. Normalization was carried out on the final prediction by mean centering and standard deviation scaling to compensate the prediction scales.

*6.3.1 Arousal and valence systems*

Regarding the arousal and valence prediction systems, an 8-mixture GMR with full covariance was used. *Principal Component Analysis* (PCA) was applied to reduce the original feature dimensionality of eGeMAPS and BoAW to 40 and 50 respectively (with ~93% and ~35% data variance preserved). Delta features and delta labels were calculated and concatenated with original features and labels to incorporate dynamic information, as per [31]. The reason for dimension reduction was to preserves enough feature variability while alleviating the need to have a large number of parameters for GMM.

The GPR subsystem was applied to the eGeMAPS features with 4s window, using PCA to reduce the dimensionality to 40, a time kernel lengthscale of 1.13 seconds for arousal and 4.54 for valence, and a presumed input noise standard deviation of 0.010 and 0.007, respectively. The 80 feature space kernel parameters were obtained by likelihood maximization on the training data, and all others system parameters were optimized using the Spearmint optimizer [54], based on the maximization of the average CCC on a 2-fold cross validation over the training data.

RVM was applied to BoAW features, Phonetic-aware (PA) features and two sets of video features (Fig. 2). For PA features, we used 59-dimensional Hungarian phoneme sets applied on 23-dimensional eGeMAPS LLDs, ending up with a set of 1357-dimensional features. The phonetic awareness parameter $\alpha$ was set to 0.2 for both arousal and valence, optimized again based on 4-fold cross validation. The iteration number of RVM for these four sets of features was optimized experimentally within the range of [10,100] with a step of 10. When performing OA fusion, the OA window size was optimized from among [10, 20, 30, 50, 80, 100, 150, 200].

*6.3.2 Liking systems*

For liking systems, besides the bag-of-text-words (BoTW) features, word affect features comprising 32-dimensional ANEW features, 20-dimensional EmoLex features and 30-dimensional SENTIC features, were investigated. RVM was used as a backend regression model, with the same iteration range of [10, 100]. Fusion of these two feature sets was carried out at both feature-level and decision-level. For feature-level fusion, word affect features concatenated with the BoTW were modelled by a single RVM. OA regression is utilized to fuse predictions in decision-level. The optimal OA window for decision-level fusion was selected as 6 sec.

# 7 DEVELOPMENT RESULTS

## 7.1 Depression Systems

7.1.1 *Text-based Feature Performance*

The first depression experiment was to evaluate the effectiveness of the proposed text-based features for predicting PHQ-8 scores. The evaluated feature sets include linguistic (i.e. siNLP, TAALES), speech auxiliary behavior, and word affect features (i.e. ANEW, EmoLex, SenticNet, and Lasswell), as seen in Fig. 4.

Compared to development baseline performance, the siNLP and speech behavior features have similar results, whilst the word affect features yielded considerably better performance, especially for the ANEW features (5.93). In addition, by using ANEW features with phrase selection (i.e. valence features of 0 value were removed), RMSE was further reduced to 5.73. The TAALES features had an RMSE as high as 9.87, which is presumably because the high dimensionality (405) of this feature set incurred over fitting. Note that the parameter was selected from 3-fold cross validation, and may not generalize well for high-dimensional features.

Interestingly, the comparison among the text-based features suggested that all four sets of word affect features considerably outperform both linguistic and speech behavior features, achieving RMSEs around or below 6.0. This is sensible and somewhat expected, since the word affect features, which potentially interpret a person's emotions such as arousal and valence, can capture the differences in emotion occurrence between healthy and depressed speakers. The emotion manifested in words can be captured by the word affect features.

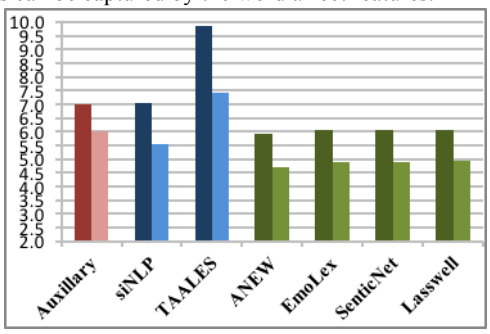


**Figure 4: Comparative SVR PHQ-8 development set prediction results using different text-based feature types: speech behavior (red), linguistic (blue), and word affect (green); RMSE (darker shade) and MAE (lighter shade)**

*7.1.2 Fusion of audio, video and word affect features*

The second experiment was to explore whether the GSR system can be effectively applied to build audio, video, and text-based individual systems, as well as for fusing different modalities. Not only can GSR, a technique based on comparisons between high severity and low severity, potentially be applied on different modalities [53], but it can also offer a good opportunity for fusing them. This is because different modalities, which are likely to have distinct features characteristics, can be mapped into a likelihood space via the high-low comparisons.

**Table 1: Performance Comparison of GSR systems on Development data set**

| | RMSE | MAE |
|---|---|---|
| (MFCC+Δ+ΔΔ) [Audio] | 6.044 | 4.835 |
| AUs [Video] | 6.673 | 5.335 |
| ANEW [Text] | 6.696 | 5.479 |
| GSR based fusion | 6.078 | 4.800 |
| Decision level fusion | 5.930 | 4.728 |

Table 1 shows the experimental results for feature sets of different modalities into the GSR system. We only selected the

best performing word affect features, i.e. ANEW, for the text-based GSR. For individual systems, the audio-based GSR system achieved strong performance compared with the video, and text-based GSR systems. One plausible reason behind this is that the 39-dimensional MFCC, which are often effectively modelled by GMM, can more accurately capture the high-low comparisons via Gaussian modelling, thereby having lower RMSE.

Apart from individual GSR systems, GSR was also used for fusion of audio, video, and text-based systems. In GSR-based fusion, unimodal LMLR vectors from each modality were concatenated to train a SVR, whereas in the decision-level fusion, predicted scores from audio, video, text-based GSR systems were concatenated to train a SVR to predict PHQ scores. However, fusion results indicate that only marginal improvement can be attained via decision-level fusion, but combining MLML scores did not aid depression prediction. Despite this, it should be noted that the RMSEs of the GSR-based systems are better than the baseline of a direct use of regression models applied on MFCCs, AUs and ANEW features (not shown here).

## 7.2 Arousal and Valence

*7.2.1 Subsystem performance*

For predicting arousal and valence, we firstly investigated performance of features from different modalities, including audio, video, and text using three types of regression models. The selection of backend regression models was empirical.

Among all the 7 subsystems shown in Table 2, the Video1+RVM performed the best, achieving 0.518 for arousal and 0.583 for valence. Note that these results have already considerably outperformed the multimodal baseline performance on the development set, let alone the video baseline. Interestingly, the BoTW+RVM also achieved very good performance, suggesting the effectiveness of text-based features for arousal and valence prediction, and the effectiveness of RVM dealing with high-dimensional feature sets for emotion prediction. The two acoustic systems of eGeMAPS+GMR and BoAW+GMR showed similar performance for arousal, while eGeMAPS+GMR was outperformed by BoAW+GMR for valence. This may indicate that BoW method applied to LLDs is capable to capture a more informative representation of emotion information for valence. Although the BoW representation in general provided reasonably good performance for both arousal and valence prediction, they are very high dimensional, which has higher risk of over fitting. An unexpected low performance was obtained by eGeMAPS+GPR and PA+RVM for arousal and valence while they achieved relative good performance when using cross validation, which is possibly owing to the generalization issue.

**Table 2: Subsystem performances with different features and back-ends**

| Features | Back-end | Arousal | Valence |
|---|---|---|---|
| eGeMAPS | GMR | 0.454 | 0.446 |
| BoAW | | 0.451 | 0.515 |
| eGeMAPS | GPR | 0.315 | 0.368 |
| PA | RVM | 0.400 | 0.362 |
| BoTW | | 0.441 | 0.499 |
| Norm Facial | | 0.518 | 0.583 |
| BoVW | | 0.397 | 0.422 |

*7.2.2 Probabilistic fusion incorporating uncertainty*

OA fusion was applied to a subset of the 7 subsystems or the entire subsystems, containing 4 audio subsystems, 2 video subsystems and 1 text subsystem. Probabilistic predictions were incorporated within the OA framework as presented in Section 5.2. A variety of combinations among 7 subsystems was investigated and the optimal sets of 7 subsystems were determined primarily according to the fusion performance of the probabilistic predictions. A comparison of the similar system configurations without and with uncertainty was also demonstrated in Table 3.

The first three sets of results is for multimodal fusion of 1) audio-only systems (i.e. eGeMAPS+GMR and BoAW+GMR), 2) audio+text systems (i.e. eGeMAPS+GMR, BoAW+GMR, eGeMAPS+GPR, PA+RVM, and BoTW+RVM), and 3) audio+text+video systems (i.e. all the seven subsystems). A comparison among the three systems suggests that including additional modalities improved CCC performance for both arousal and valence, achieving 0.620 and 0.682 respectively when fusing all subsystems. However, subsystems within the same modality may carry similar information, which could be somewhat redundant during fusion, especially for the very high dimensional BoAW and BoVW features. For this reason, we evaluated a system where a set of subsystems was selected: 4) PA+RVM, BoTW+RVM, and NormalizedFace+RVM and 5) eGeMAPS+GMR + 4), owing to its good performance and capability to incorporate the uncertainty predictions. Basically, dropping subsystems did improve the CCC of arousal from 0.620 to 0.672; however, this did not aid valence prediction, for which the performance degraded from 0.682 to 0.605.

Of particular interest is that consistent improvements were observed by incorporating prediction uncertainty across different system configurations for predicting arousal and valence, with relative improvement for arousal between 0.8% and 2.3%. Overall, the results in Table 3 considerably outperformed the multimodal baseline for both arousal and valence on the development set.

**Table 3: Fusion performance without and with uncertainty for arousal and valence in terms of CCC; numbers within brackets indicate the number of subsystems used.**

| | | Without | With |
|---|---|---|---|
| Arousal | Audio-only (2) | 0.490 | 0.494 |
| | Audio+Text (5) | 0.551 | 0.560 |
| | Audio+Text+Video(7) | 0.609 | 0.620 |
| | Audio+Text+Video (3) | 0.657 | N/A |
| | Audio+Text+Video (4) | 0.657 | **0.672** |
| Valence | Audio-only(2) | 0.500 | 0.507 |
| | Audio+Text(5) | 0.597 | 0.597 |
| | Audio+Text+Video(7) | 0.676 | **0.682** |
| | Audio+Text+Video (3) | 0.602 | N/A |
| | Audio+Text+Video (4) | 0.605 | 0.605 |

## 7.3 Liking

*7.3.1 Word affect features*

Three sets of word affect features were analyzed for liking prediction. A moving average filter with optimal length tuned within [10,100] was applied to each feature dimension to smooth

the word affect features, in order to relax the sparsity of phrase-level features. Table 4 shows the performance for each set of word affect features, as well as feature-level fusion of them. Individual systems showed different degrees of effectiveness, but were still outperformed by the BoTW features (0.314). A feature level fusion of the three feature sets yielded slight improvement.

**Table 4**: **Comparison of word affect features for liking**

| Features | CCC |
|---|---|
| ANEW | 0.160 |
| EmoLex | 0.085 |
| SenticNet | 0.150 |
| ANEW+SenticNet | 0.173 |
| ANEW+EmoLex+SenticNet | 0.172 |

*7.3.2 Fusion*

Furthermore, the proposed word affect features and the BoTW features were fused together both at feature-level and decision-level with RVM as back-end regression model. It can be seen that both fusion methods led to improvement in CCC, 0.354 and 0.352, which outperform the liking baseline on the development set.

**Table 5: Fusion performance for liking in terms of CCC**

| | CCC |
|---|---|
| (Word affect + BoTW) -RVM | **0.354** |
| Word affect+RVM & BoTW +RVM | **0.352** |

# 8 AVEC CHALLENGE RESULTS

## 8.1 Depression Systems

We submitted five systems as entries for the depression sub-challenge. It can be seen that all our submitted systems outperform baseline performance in RMSE, with the best RMSE=6.02 achieved by ANEW with phrase selection. Phrases that attained valence ratings 0 were removed, assuming there was no emotion-related information contained. This is surprising since it is the simplest system with only 32-dimensional word affect features extracted from approximately one third of the entire phrases. The results herein strongly suggest that approaching emotion-related information from transcripts is effective for depression prediction. However, unfortunately, as the systems getting more complex by including more information, either different feature sets or systems via fusion, the performance was surprisingly degraded, which is presumably because of over fitting; An insight from the test submissions is that simple systems with informative features is favored for building a robust, well-generalized depression prediction system. In this regard, the word affect features have experimentally proved to be very effective for predicting depression severity with good generalization.

**Table 6:** : **Comparison of AVEC 2017 test set for depression**

| TEST SUBMISSIONS | RMSE | MAE |
|---|---|---|
| Baseline: Audio | 7.78 | 5.72 |
| Baseline Video | 6.97 | 6.12 |
| Baseline Audio + Video | 7.05 | 5.66 |
| ANEW | **6.34** | 5.16 |
| ANEW(with phrase selection) | **6.02** | 4.98 |
| ANEW+EmoLext+SenticNet+Lasswell | 6.51 | 5.36 |
| ANEW + Audio-based GSR | 6.39 | 5.06 |
| ANEW + Audio-GSR + Video-GSR | 6.52 | 5.30 |

## 8.2 Emotion Systems

Five emotion prediction systems for arousal/valence and four liking prediction systems were submitted as emotion sub-challenge entries, shown in Table 7 and 8. Dropping subsystems within the same modality yielded the best CCC of 0.523 for arousal and 0.548 for valence, with relative improvements of 39.5% and 17.6% over the baseline on the test set.

**Table 7**: **Comparison of AVEC 2017 test set for arousal and valence**

| Features | Arousal | Valence |
|---|---|---|
| Baseline | 0.375 | 0.466 |
| Audio-only(2) | 0.344 | 0.346 |
| Audio+Text(5) | 0.377 | 0.383 |
| Audio+Text+Video(7) | 0.410 | 0.507 |
| Audio+Text+Video (3) | **0.523** | 0.540 |
| Audio+Text+Video (4) | 0.448 | **0.548** |

Word affect features as our 1st submission for liking on their own could not outperform the baseline (0.246), but still showed potential with a CCC of 0.160. Further, the proposed word affect features were fused with the BoTW features as our 2nd and 3rd submission, achieving a CCC as high as 0.318, which is 29.3% relative improvement over the baseline. Our 4th submission was a random draw from a Gaussian process with similar smoothness characteristics to the training labels, aiming to test the reliability of the liking ground truth, which did not show good performance.

**Table 8**: **Comparison of AVEC 2017 test set for liking**

| | Liking |
|---|---|
| Baseline | 0.048(0.246) |
| Word affect-RVM | 0.160 |
| (Word affect +BoTW) -RVM | 0.285 |
| Word affect-RVM + BoTW-RVM | **0.318** |
| Random Draw | 0.018 |

# 9 CONCLUSION

Our systems primarily focused on word affect features for both sub-challenges and fusion of probabilistic predictions for arousal and valence. For the depression sub-challenge, text-based features, especially word affect features, provided significant improvements of 13.6% in RMSE over the baseline on the test set. However, incorporating the audio-based and video-based GSR system did not aid depression prediction.

For the emotion sub-challenge, we fused subsystems from different modalities within the OA-RVM framework. In particular, fusion of prediction uncertainty was found to improve system performance. The proposed word affect features showed great potential for predicting liking, and considerable improvement was attained when they are fused with the BoTW features. Our best proposed emotion prediction systems provided huge improvements over the baseline, which are 39.5%, 17.6%, and 29.3% for arousal, valence, and liking.

Future work involves exploration of a larger set of word affect features. Also, the probabilistic predictions will be further analyzed to help determine a better way to aid emotion prediction.

# REFERENCES


[1] S. L. Burcusa and W. G. Iacono, "Risk for recurrence in depression," Clinical psychology review, vol. 27, no. 8, pp. 959-985, 2007.

[2] R. Frankham et al., "Predicting the probability of outbreeding depression," Conservation Biology, vol. 25, no. 3, pp. 465-475, 2011.

[3] W. H. Organization, The World Health Report 2001: Mental health: new understanding, new hope. World Health Organization, 2001.

[4] W. H. Organization, "Depression and other common mental disorders: global health estimates," 2017.

[5] R. Rickenberg and B. Reeves, "The effects of animated characters on anxiety, task performance, and evaluations of user interfaces," in Proceedings of the SIGCHI conference on Human Factors in Computing Systems, 2000, pp. 49-56: ACM.

[6] O. Grynszpan, J.-C. Martin, and J. Nadel, "Human computer interfaces for autism: assessing the influence of task assignment and output modalities," in CHI'05 Extended Abstracts on Human Factors in Computing Systems, 2005, pp. 1419-1422: ACM.

[7] N. Cummins, V. Sethu, J. Epps, S. Schnieder, and J. Krajewski, "Analysis of acoustic space variability in speech affected by depression," Speech Communication, vol. 75, pp. 27-49, 2015.

[8] D. Matsumoto, T. Kudoh, K. Scherer, and H. Wallbott, "Antecedents of and reactions to emotions in the United States and Japan," Journal of Cross-Cultural Psychology, vol. 19, no. 3, pp. 267-286, 1988.

[9] C. J. O'Byrne and S. Flint, "High-resolution sequence stratigraphy of Cretaceous shallow marine sandstones, Book Cliffs outcrop, Utah, USA-application to reservoir modelling," First Break, vol. 11, no. 10, pp. 445-459, 1993.

[10] J. K. Darby, N. Simmons, and P. A. Berger, "Speech and voice parameters of depression: A pilot study," Journal of Communication Disorders, vol. 17, no. 2, pp. 75-85, 1984.

[11] D. Bennabi, P. Vandel, C. Papaxanthis, T. Pozzo, and E. Haffen, "Psychomotor retardation in depression: a systematic review of diagnostic, pathophysiologic, and therapeutic implications," BioMed research international, vol. 2013, 2013.

[12] W. Bucci and N. Freedman, "The language of depression," Bulletin of the Menninger Clinic, vol. 45, no. 4, p. 334, 1981.

[13] T. K. Witte, K. A. Timmons, E. Fink, A. R. Smith, and T. E. Joiner, "Do major depressive disorder and dysthymic disorder confer differential risk for suicide?," Journal of affective disorders, vol. 115, no. 1, pp. 69-78, 2009.

[14] A. Feinstein, S. Magalhaes, J.-F. Richard, B. Audet, and C. Moore, "The link between multiple sclerosis and depression," Nature Reviews Neurology, vol. 10, no. 9, pp. 507-517, 2014.

[15] J. D. Williamson et al., "Intensive vs standard blood pressure control and cardiovascular disease outcomes in adults aged≥ 75 years: a randomized clinical trial," Jama, vol. 315, no. 24, pp. 2673-2682, 2016.

[16] K. Kyle and S. A. Crossley, "Automatically assessing lexical sophistication: Indices, tools, findings, and application," Tesol Quarterly, vol. 49, no. 4, pp. 757-786, 2015.

[17] N. J. Andreasen and B. Pfohl, "Linguistic analysis of speech in affective disorders," Archives of General Psychiatry, vol. 33, no. 11, pp. 1361-1367, 1976.

[18] S. Rude, E.-M. Gortner, and J. Pennebaker, "Language use of depressed and depression-vulnerable college students," Cognition & Emotion, vol. 18, no. 8, pp. 1121-1133, 2004.

[19] K. Collier, B. Bickel, C. P. van Schaik, M. B. Manser, and S. W. Townsend, "Language evolution: syntax before phonology?," in Proc. R. Soc. B, 2014, vol. 281, no. 1788, p. 20140263: The Royal Society.

[20] B. Ojamaa, P. K. Jokinen, and K. Muischenk, "Sentiment analysis on conversational texts," in Proceedings of the 20th Nordic Conference of Computational Linguistics, NODALIDA 2015, May 11-13, 2015, Vilnius, Lithuania, 2015, no. 109, pp. 233-237: Linköping University Electronic Press.

[21] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting Depression via Social Media," ICWSM, vol. 13, pp. 1-10, 2013.

[22] Z. Huang et al., "An investigation of annotation delay compensation and output-associative fusion for multimodal continuous emotion prediction," in Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge, 2015, pp. 41-48: ACM.

[23] B. Stasak and J. Epps, "Elicitation design for acoustic depression classification: an investigation of articulation effort, linguistic complexity, and word affect," presented at the InterSpeech 17, Stockholm – Sweden, accepted, 2017., 2017.

[24] F. Eyben et al., "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," IEEE Transactions on Affective Computing, vol. 7, no. 2, pp. 190-202, 2016.

[25] F. Ringeval and B. Schuller, "AVEC2017-Real-life Depression, and Affect Recognition Workshop and Chanllenge," presented at the in Proceedings of the 7th International Workshop on Audio/Visual Emotion Challenge,2017,ACM, 2017.

[26] Z. Huang and J. Epps, "A PLLR and Multi-stage Staircase Regression Framework for Speech-based Emotion Prediction," 2017.

[27] E. Mower et al., "Interpreting ambiguous emotional expressions," in Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on, 2009, pp. 1-8: IEEE.

[28] T. Dang, V. Sethu, and J. Epps, "An Investigation of Emotion Prediction Uncertainty Using Gaussian Mixture

Regression," presented at the Interspeecj 17, Stockholm,Sweden, 2017.

[29] M. Grimm, K. Kroschel, and S. Narayanan, "Support vector regression for automatic recognition of spontaneous emotions in speech," in Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on, 2007, vol. 4, pp. IV-1085-IV-1088: IEEE.

[30] M. Wöllmer, M. Kaiser, F. Eyben, B. Schuller, and G. Rigoll, "LSTM-modeling of continuous emotions in an audiovisual affect recognition framework," Image and Vision Computing, vol. 31, no. 2, pp. 153-163, 2013.

[31] A. Metallinou, A. Katsamanis, Y. Wang, and S. Narayanan, "Tracking changes in continuous emotion states using body language and prosodic cues," in Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, 2011, pp. 2288-2291: IEEE.

[32] J. M. Wang, D. J. Fleet, and A. Hertzmann, "Gaussian process dynamical models for human motion," IEEE transactions on pattern analysis and machine intelligence, vol. 30, no. 2, pp. 283-298, 2008.

[33] M. A. Nicolaou, H. Gunes, and M. Pantic, "Output-associative rvm regression for dimensional and continuous emotion prediction," Image and Vision Computing, vol. 30, no. 3, pp. 186-196, 2012.

[34] Z. Huang et al., "Staircase Regression in OA RVM, Data Selection and Gender Dependency in AVEC 2016," in Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, 2016, pp. 19-26: ACM.

[35] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP—A collaborative voice analysis repository for speech technologies," in Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, 2014, pp. 960-964: IEEE.

[36] M. Schmitt, F. Ringeval, and B. W. Schuller, "At the Border of Acoustics and Linguistics: Bag-of-Audio-Words for the Recognition of Emotions in Speech," in INTERSPEECH, 2016, pp. 495-499.

[37] K. Kyle. http://www.kristopherkyle.com/.

[38] S. A. Crossley, L. K. Allen, K. Kyle, and D. S. McNamara, "Analyzing discourse processing using a simple natural language processing tool," Discourse Processes, vol. 51, no. 5-6, pp. 511-534, 2014.

[39] S. A. Crossley, K. Kyle, and D. S. McNamara, "Sentiment Analysis and Social Cognition Engine (SEANCE): An automatic tool for sentiment, social cognition, and social-order analysis," Behavior research methods, vol. 49, no. 3, pp. 803-821, 2017.

[40] M. M. Bradley and P. J. Lang, "Affective norms for English words (ANEW): Instruction manual and affective ratings," Technical report C-1, the center for research in psychophysiology, University of Florida1999.

[41] S. M. Mohammad and P. D. Turney, "Crowdsourcing a word–emotion association lexicon," Computational Intelligence, vol. 29, no. 3, pp. 436-465, 2013.

[42] E. Cambria, R. Speer, C. Havasi, and A. Hussain, "SenticNet: A Publicly Available Semantic Resource for Opinion Mining," in AAAI fall symposium: commonsense knowledge, 2010, vol. 10, no. 0.

[43] H. D. Lasswell and J. Z. Namenwirth, "The Lasswell value dictionary," New Haven, 1969.

[44] J. R. L.-B. Bernard and J. R. L.-B. Bernard, The Macquarie Thesaurus. Macquarie, 1986.

[45] T. Brants and A. Franz, "Web 1t 5-gram version 1 (2006)," Linguistic Data Consortium, Philadelphia.

[46] P. J. Stone, D. C. Dunphy, and M. S. Smith, "The general inquirer: A computer approach to content analysis," 1966.

[47] J. Z. Namenwirth and R. P. Weber, Dynamics of culture. Routledge, 2016.

[48] H. Zhaocheng and J. Epps, "An Investigation of Partition-based and Phonetically-aware Acoustic Features for Continuous Emotion Prediction from Speech," Transactions on Affective Computing, vol. Under review.

[49] D. K. Duvenaud, H. Nickisch, and C. E. Rasmussen, "Additive Gaussian processes," in Advances in neural information processing systems, 2011, pp. 226-234.

[50] J. Gratch et al., "The Distress Analysis Interview Corpus of human and computer interviews," in LREC, 2014, pp. 3123-3128.

[51] K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. Williams, J. T. Berry, and A. H. Mokdad, "The PHQ-8 as a measure of current depression in the general population," Journal of affective disorders, vol. 114, no. 1, pp. 163-173, 2009.

[52] B. Schuller, J.-G. Ganascia, and L. Devillers, "Multimodal Sentiment Analysis in the Wild: Ethical considerations on Data Collection, Annotation, and Exploitation," in Actes du Workshop on Ethics In Corpus Collection, Annotation & Application (ETHI-CA2), LREC, Portoroz, Slovénie, 2016.

[53] J. R. Williamson, T. F. Quatieri, B. S. Helfer, R. Horwitz, B. Yu, and D. D. Mehta, "Vocal biomarkers of depression based on motor incoordination," in Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge, 2013, pp. 41-48: ACM.

[54] K. Eggensperger et al., "Towards an empirical foundation for assessing bayesian optimization of hyperparameters," in NIPS workshop on Bayesian Optimization in Theory and Practice, 2013, vol. 10.



View publication stats