INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Capturing Mismatch between Textual and Acoustic Emotion Expressions for Mood Identification in Bipolar Disorder

Minxue Niu, Amrit Romana, Mimansa Jaiswal, Melvin McInnis, Emily Mower Provost

University of Michigan, USA

{sandymn, aromana, mimansa, mmcinnis, emilykmp}@umich.edu

Abstract

Emotion is a complex behavioral phenomenon, which is expressed and perceived through various modalities, such as language, vocal and facial expressions. Psychiatric research has suggested that the lack of emotional alignment between modalities is a symptom of emotion disorders. In this work, we quantify the mismatch between emotion expressed through language and acoustics, which we refer to as Emotional MisMatch (EMM), as an intermediate step for mood identification. We use a longitudinal dataset collected from people with Bipolar Disorder (BP) and show that symptomatic mood episodes show significantly more EMM, compared to euthymic moods. We propose a fully automatic mood identification pipeline with automatic speech transcription, emotion recognition, and EMM feature extraction. We find that EMM features, although smaller in size, outperform a language-based baseline, and consistently provide improvement when combined with language and/or raw emotion features on mood classification.

Index Terms: speech, emotion, mood, bipolar disorder

1. Introduction

Providing adequate care for individuals living with mental health conditions represents one of the major challenges in society. Mood disorders are often chronic and recurring, and require ongoing monitoring with consistent access to specialty mental health care providers. There is a substantial gap in the availability of such providers and the demand for service [1]. This motivates the need to develop tools to monitor mood and prioritize individuals for care.

Researchers have proposed automatic mood tracking techniques using various types of data, but the data collection process usually requires considerable effort [2]. Speech and language data have been widely used, the benefit being that speech can be recorded in natural environments and has shown promise in automated mood identification [3]. Commonly used linguistic feature sets include frequency counts on categories of words that encode linguistic styles and semantic content [4] (e.g., Linguistic Inquiry and Word Count (LIWC) [5]). Commonly used acoustic feature sets include aspects that capture the frequency content or prosodic patterns of an individual's vocalizations [6].

Another line of research involves survey methodologies (e.g., Ecological Momentary Assessment [7]), where participants complete surveys in which they report their feelings and behaviour in their daily lives. Research has demonstrated the close connection between emotion variations and mood change, supporting the hypothesis that mood identification models benefit from the explicit modeling of emotion characteristics [8]. However, the use of these tools requires that individuals take actions outside of their day-to-day life, which is difficult to sustain in the long term. There have also been works where emotion recognition models were applied to obtain emotion ratings on speech or text data for mood prediction, eliminating the need for surveys [9, 10]. Those works mostly rely on model training processes to learn the relationship between emotion and mood variations, seldom incorporating domain knowledge obtained from clinical research.

In this work, we aim to bring together the benefits of emotion recognition models and clinical insights. We propose a set of emotion-centered features inspired by the phenomena of Affect Blunting (AB) and Affect Exaggeration (AE). AB refers to when individuals show reduced emotional expressivity despite strong self-reported emotion and is a symptom associated with schizophrenia and depression [11]. Subjects suffering from AB can fail to show normal vocal emphasis patterns [12], where the text content may be reflective of their emotions but the acoustic content is not. AE, the symptom where individuals show exaggerated affect, is a typical symptom in mania [13]. In this work, we hypothesize that including EMM features in mood prediction models will bring extra improvement to their performance on top of raw emotion ratings.

We study EMM in a longitudinal dataset collected from individuals with Bipolar Disorder (BP) [14]. We use existing acoustic emotion labels and crowdsource text emotion annotations on a consented and transcribed subset of the dataset [9]. Speech segments from depressed or manic states display statistically significantly more EMM, compared to those from euthymic states. Further, we find that emotions measured from acoustics are more negative and less energetic than emotions measured from text in depressed speech, supporting clinical findings of affect blunting. We then explore how EMM can be incorporated into mood recognition systems. We define a set of EMM features over emotion scores, which aim to capture the dynamic interactions between acoustic and text emotions. We propose a fully automatic pipeline composed of speech recognition to generate transcripts and separate acoustic and text emotion recognition models to automate the extraction of EMM features. We find that with a simple linear model, EMM features outperform emotion-only features and a state-of-the-art language-feature baseline in multiple mood prediction tasks. Further, we find that they contribute to better performance when combined with baseline features. Our results show EMM contains strong and consistent signals of mood change. These informative and interpretable features have great potential to be used in mood tracking applications.

2. PRIORI dataset

The Predicting Individual Outcomes for Rapid Intervention (PRIORI) Dataset is a longitudinal collection from individuals

10.21437/Interspeech.2023-1990

Table 1: *Definition of mood classes.*

Mood	Scale Ratings
Euthymia	HDRS<=7 and YMRS <=9
Sub-clinical Depression	YMRS<=9 and 8<=HDRS<=16
Clinical Depression	HDRS>=17
Sub-clinical Mania	HDRS<=7 and 10<=YMRS<=20
Clinical Mania	YMRS>=21

with BP [14]. The dataset consists of smartphone conversations made or received by study participants over the course of six to twelve months. Calls were recorded using a secure app installed on participants' smartphones, capturing the participant's side of the conversation. Transcripts of the recordings were obtained with Microsoft Azure speech-to-text transcription service.

There are two types of calls in the dataset: 1) weekly assessment calls, in which clinicians conduct mood assessment interviews with the participant and 2) personal calls, which include all other calls. Prior work has shown that mood signals are more clear in the assessment calls than personal calls [9], and thus in this paper we focus on assessment calls.

Each assessment call contains two structured interviews using the Hamilton Depression Rating Scale (HDRS) [15] and the Young Mania Rating Scale (YMRS) [16]. We define five mood classes based on the HDRS and YMRS scores following previous work [17, 18] (Table 1). We use the terms "depressed" or "manic" mood to mean the combination of both the sub-clinical and clinical classes. We apply the same exclusion criterion as in previous work [19], which yielded a final dataset with 1022 calls from 47 subjects, with an average of 82.4 segments (continuous speech separated by silence) per call. The average length of a segment is 6.25 seconds, with a 4.24 seconds std and minimum length of 3 seconds, as in [9]. The distribution of mood labels is 559 euthymia, 389 sub-clinical depression, 129 clinical depression, 57 sub-clinical mania and 22 clinical mania. We measure performance on HDRS score regression, depression severity classification (Euthymic vs. Sub-clinical depression vs. Clinical depression) and 3-way mood classification (Euthymic vs. depression vs. mania). We don't run regression on YMRS scores due to insufficient number of manic samples.

2.1. Emotion labels and crowdsourcing

The PRIORI-Emotion Dataset is a subset of PRIORI that contains data from 19 of the subjects who consented to have researchers listening to their phone call recordings [9]. We build emotion recognition models with this subset.

Acoustic Emotion Labels. Previous work has obtained acoustic emotion annotations assessed on a 9-point Likert scale on two dimensions [20]: valence (1 very negative, 9 very positive) and activation (1 very subdued, 9 very energized). The annotators were trained to rate solely based on the acoustic aspects of the recordings, ingnoring their lecical content.

Text Emotion Labels. We obtained text emotion annotations with transcripts using the Amazon Mechanical Turk (AMT) crowdsourcing platform. All experiments were approved by University of Michigan Institutional Review Board. We asked annotators to read segment transcripts and select the valence and activation labels that best fit. We again used 9-point Likert scales to evaluate the valence and activation, as in the acoustic annotations. We required AMT workers to be in the US and to have a more than 98% history approval rate. Before the annotation started, workers read a short instruction note explaining valence and activation, and took a brief qualification test to ensure they understood the concepts. Each sample was assigned to three different workers.

We identified unreliable AMT workers by calculating the percentage of "outliers" in their annotation [21]. For each segment, we compared the annotation from the target worker ann_{worker} with the other two annotations ann_1 and ann_2. If $\min(|ann_{worker} - ann_1|, |ann_{worker} - ann_2|) > \max(2, |ann_1 - ann_2|)$, we consider ann_{worker} an outlier. This resulted in the removal of 190 annotations from 17 workers (out of 173) who had a >15% outlier rate, one standard deviation higher than the mean outlier rate. Our annotated subset has 8,033 segments with an average of 2.97 annotations per segment. We use the averaged score across annotators as the final emotion label.

3. Methods

3.1. Emotional MisMatch (EMM)

3.1.1. Quantifying EMM from emotion scores

We define segment-level EMM as the difference between acoustic and text emotion scores. We define a set of statistical descriptors on all segments from a call as the call-level EMM features.

Formally, we consider n segments in a call, $i \in \{0, n-1\}$, each associated with four emotion scores: activation and valence labels on both modalities. We use *a* and *v* for activation and valence, and *ac* and *tt* for the acoustic and text modalities, respectively. Then, a_i^{ac} represents the acoustic activation rating on the i^{th} segment, and v_i^{tt}, the textual valence rating on the i^{th} segment. We use *m* to denote the mismatch between acoustic and textual emotion, i.e., $m_i = (v_i^{ac} - v_i^{tt}, a_i^{ac} - a_i^{tt})$. We then define the EMM feature set with the following components:

Mismatch statistics (25-dim). We calculate 5 statistics (min, median, max, mean, and standard deviation) over the valence mismatch $v^{ac} - v^{tt}$, its absolute value $|v^{ac} - v^{tt}|$, the activation mismatch $a^{ac} - a^{tt}$, its absolute value $|a^{ac} - a^{tt}|$, and the mismatch distance $\|m\|$.

Covariance (4-dim). To measure the extent to which text and acoustic emotion vary together, we calculate covariance on the emotion ratings (i.e., lists of the emotion scores over time) for each call. Specifically, we calculate the covariance between v^{tt} and v^{ac}, a^{tt} and a^{ac}, v^{tt} and a^{ac}, and a^{tt} and v^{ac}.

Interaction with text emotion (20-dim). We hypothesize that the same mismatch values can indicate very different symptoms when the emotion content differs. For example, a person sounding neutral when recalling an exciting experience could convey different information than when sounding sad talking about emotionally neutral content, although they could have the same negativity $v^{ac} - v^{tt}$. To capture this difference, we add 5 statistics on the interaction terms between the emotion and EMM features: $v^{tt} \times (v^{ac} - v^{tt})$, $a^{tt} \times (v^{ac} - v^{tt})$, $v^{tt} \times (a^{ac} - a^{tt})$, and $a^{tt} \times (a^{ac} - a^{tt})$).

3.1.2. Automatic emotion recognition

In this section, we describe the emotion recognition methods we use to automatically predict the text and acoustic emotion labels from raw speech.

Text Emotion Recognition (TER). Recently, transformer-based methods have achieved state-of-the-art results. We use Bidirectional Encoder Representations from Transformers (BERT) [22] as our text feature extractor. BERT is a large pre-trained language model with outstanding performance in various natural language understanding tasks. We use the pre-

trained BERT model "bert-base-uncased" from Huggingface[1] as the embedding layer, on which we add two regression layers: one for activation and another for valence.

Acoustic Emotion Recognition (AER). Large transformer-based neural networks (such as wav2vec 2.0) have also shown competitive AER performance. However, these learned representations also embed language information [23], which is at odds with our desire to separately measure text and acoustic emotional content. In this work we use a convolutional neural network (CNN) with Mel Filterbank (MFB) features for AER. For the CNN, we adopt the same architecture that was found to be effective on these data in previous work [9]: We train a CNN that takes 40-dimensional MFB features as input. The CNN itself has with two convolutional layers, each with a kernel size of 4, and output channels of 120 and 360, respectively. The convolutional layers are followed by max pooling over time, a fully connected layer with an output size of 240, and the same two output heads, as the TER model. Each layer, other than the output layer, is followed by a ReLU activation function.

For both TER and AER, we experiment with two model architectures: the base model, which predicts activation and valence independently, and a multi-task model which jointly predicts valence and activation, with the objective function being the average error of valence and activation predictions.

Emotion Recognition Performance. We randomly split our data into nine subject-independent folds. Eight of the folds contain data from two subjects, and the last fold contains three. For each experiment, we train on seven folds, validate on one fold and test on one fold, to ensure that all data from test subjects have not been seen by the model. We measure the average performance over five random seeds and over all folds.

We train both models with an AdamW optimizer with average Root Mean Square Error (RMSE) loss on activation and valence. For TER, we train 10 epochs with a hyperparameter search across learning rates of {5e-5, 1e-5, 5e-6}, with a final selection of 5e-6. For AER, we train 15 epochs with a learning rate of 1e-4 as in [9]. We use Concordance Correlation Coefficient (CCC) as model selection criterion on the validation set.

Consistent with previous findings [24], the text modality is better at capturing valence, while acoustics perform better for activation (Table 2). We find that multi-task learning generally improves performance and reduces standard deviation, especially for the "weaker" modality. We apply the trained multi-task models to predict the valence/activation scores on segments. We use the average over the five random seeds as predicted emotion scores and use these to extract the EMM features (see Section 3.1.2).

3.2. Mood prediction models

We evaluate the potential of mismatch features for mood recognition with comparison to two baseline feature sets: 1) language and 2) emotion-only. The **language baseline** follows prior work which achieves state-of-the-art performance on mood severity prediction (HDRS) on the PRIORI dataset [19]. They proposes a set of language features that capture linguistic style, semantic content (Term Frequency Inverse Document Frequency (TF-IDF)), speech intelligibility, along with speaker timing information. We extract these features using their released code[2], which yields 111-dim language features plus tens of thousands of TF-IDF features.

The **emotion-only baseline** first extracts segment-level valence and activation for both acoustics and text and then calculates call-level statistics using the same five statistics described in the mismatch section (20-dim). We augment this with in-call emotion variance (4-dim) and the within-modality covariance (2-dim). All features are normalized by the mean and standard deviation on all euthymic segments in the training set. For both baselines, we perform feature selection with f-regression based correlation ranking. We decide the number of features to keep through cross-validation on the training set, as in [19]. We focus on three tasks: 1) linear regression to predict the HDRS scores, 2) multi-class logistic regression to predict depression severity (euthymia vs. sub-clinical-depression vs. depression), and 3) multi-class logistic regression for 3-way mood prediction (depression vs. euthymia vs. mania, Table 1). We report the mean and standard deviation of performance across subjects using leave-one-subject-out cross-validation.

Table 2: *Emotion recognition performance where tt=textual, ac=acoustic, sep=single-task, and multi=multi-task. CCC=concordance correlation coefficient and RMSE=root mean square error. Best results are shown in bold.*

Model	Activation		Valence	
	CCC	RMSE	CCC	RMSE
tt-sep	0.301 (.01)	0.242 (.00)	0.620 (.01)	**0.203 (.00)**
tt-multi	**0.328 (.00)**	**0.234 (.00)**	**0.621 (.00)**	**0.203 (.00)**
ac-sep	**0.579 (.01)**	**0.243 (.00)**	0.316 (.01)	0.248 (.00)
ac-multi	0.556 (.01)	**0.243 (.00)**	**0.397 (.02)**	**0.238 (.00)**

4. Results and analysis

4.1. EMM feature analysis

We first assess whether EMM features differ between mood classes. We run independent t-test between the mismatch distances of euthymic and symptomatic speech segments from the annotated set as a preliminary test of the utility of EMM for mood prediction. We found that speech from individuals in depressed or manic moods show significantly more mismatch than those from euthymic moods: $\mu_{euthymia}(1.04) < \mu_{depress}(1.24)$ ($p < 0.001$), $\mu_{euthymia} < \mu_{mania}(1.35)$ ($p < 0.001$). We then apply kernel density estimation on the 2-d mismatch space (m_i) for each mood state. We visualize the differences between the estimated density of symptomatic and euthymic states in Figure 1. We find that in euthymic speech, the distribution is well centered around the origin, which represents similarity between acoustic and text emotion (Figure 1a). Alternatively, depressed speech tends to sound less energetic and more negative than the conveyed content (see the red section in the lower-left area of Figure 1b), while manic speech shows more mismatch across the space (note the dispersed red, Figure 1c). Those mismatch distributions are consistent with the clinical findings on AB and AE.

4.2. Mood prediction

We measure the performance of the regression task with Pearson Correlation Coefficient (PCC) and Root Mean Squared Error (RMSE), and we measure classification performance with Un-

[1] https://huggingface.co/bert-base-uncased

[2] https://github.com/kmatton/Feature-Extraction/tree/master/text_features

Figure 1: *(a) Estimated density of emotional mismatch in valence (x-axis, $v_i^{ac} - v_i^{tt}$) and activation (y-axis, $a_i^{ac} - a_i^{tt}$) of euthymic speech. (b, c) Mismatch density differences between symptomatic (b, depressed and c, manic) and euthymic speech.*

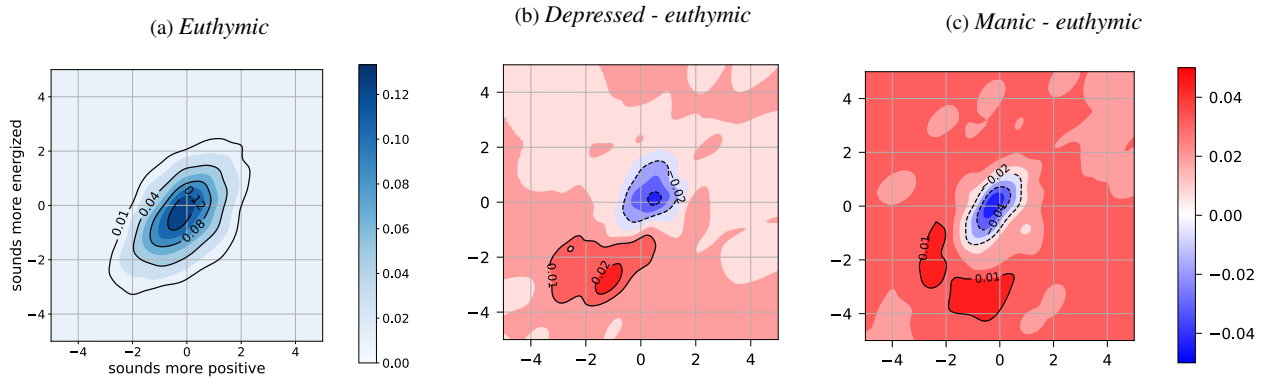

Table 3: *Mood prediction results. Lang. - language baseline, Emo. - emotion-only baseline, All - Lang. + Emo. + EMM. Best results of each task are shown in bold. Paired t-tests were performed on subject-wise performances between the language baseline and other feature sets. Significance was shown with symbols following the numbers: ˆ : 0.05<p<0.1, * : p<0.05.*

	Lang.	Emo.	EMM	Emo+EMM	All
HDRS Regression					
PCC	.38 (.31)	.41 (.22)	.47 (.25)ˆ	**.50 (.22)***	0.48 (.27)*
RMSE	5.9 (2.2)	5.6 (2.4)	5.5 (2.2)	5.5 (2.2)	**5.4 (1.9)***
Depression Severity Classification (3-class)					
UAR	.46 (.14)	.45 (.12)	.47 (.16)	.48 (.12)	**.54 (.16)***
Mood Classification (3-class)					
UAR	.43 (.05)	.42 (.07)	.42 (.07)	**.45 (.08)**	.43 (.08)

weighted Average Recall (UAR). As shown in Table 3, EMM features, when used alone, consistently outperform the baselines in the HDRS regression task: EMM features achieve 0.47 PCC on HDRS, improving over language (0.38), and also over emotion (0.41) baselines. In classification tasks, EMM features brought more improvement when used together with language and/or emotion features.

We conduct an investigation of the feature selection process with the HDRS regression model. We find that eight features are consistently selected across all cross-validation runs and these features are listed, with their coefficients, in Table 4. Of these eight features, four are mismatch features: the standard deviation of valence mismatch and three covariance values. Other features include the count of positive emotion words from LIWC, minimum text valence, and two affirmation words/bigrams from TF-IDF: "yes" and "yeah yes". Consistent with clinical findings on AB: EMM covariance features have negative coefficients. Lower covariance indicates higher mismatch, which is related to higher HDRS scores. Lower text valence score and lower LIWC positive emotion word counts are also indicators of higher depression severity. The last two TF-IDF features are associated with answers in structured clinical interviews. A positive coefficient for these features is capturing the nature of a clinical interaction where individuals answer

Table 4: *list of features that are selected across all runs, and the average of their fitted coefficients in HDRS regression.*

Feature type	Feature name	Coefficient
EMM	std($\lvert v^{ac} - v^{tt} \rvert$)	0.025
	cov(v^{tt}, v^{ac})	-0.064
	cov(v^{tt}, a^{ac})	-1.389
	cov(v^{ac}, v^{ac})	-1.134
Emotion	min(v^{tt})	-0.934
Language	(LIWC) positive emotion	-0.415
	(TFIDF) "yes"	1.200
	(TFIDF) "yeah yes"	0.341

"yes" to questions on symptoms they may be experiencing. As a result, we suspect that these features may not generalize well to natural speech [25, 19].

5. Conclusion

Inspired by research on Affect Blunting and Affect Exaggeration, we propose a set of EMM features quantifying the mismatch between text and acoustic emotion expressions. We show EMM feature extraction can be fully automated on raw speech recordings using ASR and uni-modal emotion recognition models. Symptomatic speech shows significantly more mismatch than euthymic speech. Our EMM features outperform existing language and emotion-only mood baselines, and they can also be used together with existing features to further improve performance. Our results support findings from previous work that mood recognition models benefit from learning emotion variations as an intermediate step, and can benefit from multimodal input by observing the mismatch and interactions between them. Looking forward, we will investigate how these features perform in other tasks that relate to mental health. We further plan to explore how models can be guided to attend to meaningful mismatch patterns between modalities, to promote the learning of EMM patterns.

6. Acknowledgements

This material is based in part upon work supported by the National Science Foundation (NSF IIS-RI 2006618).

7. References

[1] W. H. Organization *et al.*, "World mental health report: transforming mental health for all: executive summary," in *World mental health report: transforming mental health for all: executive summary*, 2022.

[2] G. Malhi, A. Hamilton, G. Morris, Z. Mannie, P. Das, and T. Outhred, "The promise of digital mood tracking technologies: are we heading on the right track?" *Evidence-based mental health*, vol. 20.

[3] M. Morales and R. Levitan, "Speech vs. text: A comparative analysis of features for depression detection systems," in *IEEE spoken language technology workshop (SLT)*. IEEE, 2016, pp. 136–143.

[4] J. Deng and F. Ren, "A survey of textual emotion recognition and its challenges," *IEEE Transactions on Affective Computing*, 2021.

[5] Y. Tausczik and J. Pennebaker, "The psychological meaning of words: Liwc and computerized text analysis methods," *Journal of language and social psychology*, vol. 29.

[6] O. Flanagan, A. Chan, P. Roop, F. Sundram *et al.*, "Using acoustic speech patterns from smartphones to investigate mood disorders: Scoping review," *JMIR mHealth and uHealth*, vol. 9.

[7] S. Shiffman, A. Stone, and M. Hufford, "Ecological momentary assessment," *Annu. Rev. Clin. Psychol.*, vol. 4, pp. 1–32, 2008.

[8] B. H. Morris, L. M. Bylsma, and J. Rottenberg, "Does emotion predict the course of major depressive disorder? a review of prospective studies," *British Journal of Clinical Psychology*, vol. 48.

[9] S. Khorram, M. Jaiswal, J. Gideon, M. McInnis, and E. Mower Provost, "The priori emotion dataset: Linking mood to emotion detected in-the-wild," *Interspeech*, 2018.

[10] J. Gideon, H. Schatten, M. McInnis, and E. Mower Provost, "Emotion recognition from natural phone conversations in individuals with and without recent suicidal ideation," in *Interspeech*, 2019.

[11] H. Berenbaum and T. Oltmanns, "Emotional experience and expression in schizophrenia and depression." *Journal of abnormal psychology*, vol. 101.

[12] N. Andreasen, "The scale for the assessment of negative symptoms (sans): conceptual and theoretical foundations," *The British journal of psychiatry*, vol. 155.

[13] K. M'Bailara, T. Atzeni, F. Colom, J. Swendsen, S. Gard, A. Desage, and C. Henry, "Emotional hyperreactivity as a core dimension of manic and mixed states," *Psychiatry research*, vol. 197.

[14] Z. Karam, E. Mower Provost, S. Singh, J. Montgomery, C. Archer, G. Harrington, and M. Mcinnis, "Ecologically valid long-term mood monitoring of individuals with bipolar disorder using speech," in *IEEE international conference on acoustics, speech and signal processing*, 2014, pp. 4858–4862.

[15] M. Hamilton, "A rating scale for depression," *Journal of neurology, neurosurgery, and psychiatry*, vol. 23, no. 1, p. 56, 1960.

[16] R. Young, J. Biggs, V. Ziegler, and D. Meyer, "A rating scale for mania: reliability, validity and sensitivity," *The British journal of psychiatry*, vol. 133.

[17] S. McElroy, B. Martens, R. Creech, J. Welge, L. Jefferson, A. Guerdjikova, and P. Keck Jr, "Randomized, double-blind, placebo-controlled study of divalproex extended release loading monotherapy in ambulatory bipolar spectrum disorder patients with moderate-to-severe hypomania or mild mania," *The Journal of clinical psychiatry*, vol. 71.

[18] M. Zimmerman, J. Martinez, D. Young, I. Chelminski, and K. Dalrymple, "Severity classification on the hamilton depression rating scale," *Journal of affective disorders*, vol. 150.

[19] K. Matton, M. McInnis, and E. Mower Provost, "Into the wild: Transitioning from recognizing mood in clinical interactions to personal conversations for individuals with bipolar disorder," in *Interspeech*, 2019.

[20] J. Russell, "Core affect and the psychological construction of emotion." *Psychological review*, vol. 110.

[21] S. Jagabathula, L. Subramanian, and A. Venkataraman, "Identifying unreliable and adversarial workers in crowdsourced labeling tasks," *The Journal of Machine Learning Research*, vol. 18.

[22] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv:1810.04805*, 2018.

[23] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Eyben, and B. Schuller, "Dawn of the transformer era in speech emotion recognition: closing the valence gap," *arXiv:2203.07378*, 2022.

[24] M. Perez, M. Jaiswal, M. Niu, C. Gorrostieta, M. Roddy, K. Taylor, R. Lotfian, J. Kane, and E. Mower Provost, "Mind the gap: On the value of silence representations to lexical-based speech emotion recognition," *Interspeech*.

[25] J. Gideon, E. M. Provost, and M. McInnis, "Mood state prediction from speech of varying acoustic quality for individuals with bipolar disorder," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 2359–2363.