Title

Using machine learning analyses of speech to classify levels of expressed emotion in parents of youth with mood disorders.

Authors

Weintraub, Marc J
Posta, Filippo
Arevian, Armen C
et al.

Publication Date

2021-04-01

DOI

10.1016/j.jpsychires.2021.01.019

Peer reviewed

Published in final edited form as:
J Psychiatr Res. 2021 April ; 136: 39–46. doi:10.1016/j.jpsychires.2021.01.019.

Using Machine Learning Analyses of Speech to Classify Levels of Expressed Emotion in Parents of Youth with Mood Disorders

Marc J. Weintraub[a], **Filippo Posta**[b], **Armen C. Arevian**[a], **David J. Miklowitz**[a]

[a]UCLA Semel Institute, Los Angeles, CA

[b]Estrella Mountain Community College, Avondale, AZ

Abstract

Expressed emotion (EE), a measure of attitudes among caregivers towards a patient with a psychiatric disorder, is a robust predictor of relapse across mood and psychotic disorders. Because the measurement of EE is time-intensive and costly, its use in clinical settings has been limited. In an effort to automate EE classification, we evaluated whether machine learning (ML) applied to lexical features of speech samples can accurately categorize parents as high or low in EE or in its subtypes (criticism, overinvolvement, and warmth). The sample was 123 parents of youth who had active mood symptoms and a family history of bipolar disorder. Using ML algorithms, we achieved 75.2–81.8% accuracy (sensitivities of ~0.7 and specificities of ~0.8) in classifying parents as high or low in EE and EE subtypes. Further, machine-derived EE classifications' relationships with mood symptoms, parental distress, and family conflict paralleled observer-rated EE classifications' relationships with the same variables. Of note, criticism related to greater manic severity, parental distress, and family conflict. Study findings indicate that EE classification can be automated through lexical analysis and suggest potential for facilitating larger-scale applications in clinical settings. The results also provide initial indications of the digital phenotypes that underlie EE and its subtypes.

Keywords

digital phenotype; mood disorder; depression; criticism; overinvolvement; warmth

Corresponding Author: Marc J. Weintraub, Ph.D., UCLA Semel Institute, Department of Psychiatry, 760 Westwood Plaza, A7-370, Los Angeles, CA 90095, mjweintraub@mednet.ucla.edu.

Contributions:
Drs. Weintraub, Arevian, & Miklowitz designed the study and oversaw the study procedures. Dr. Miklowitz was in charge of the study administration. Dr. Posta undertook the statistical analyses. Dr. Weintraub wrote the first draft of the manuscript, and all authors contributed to and have approved the final manuscript submitted here.

Declaration of competing interest:
The authors declare they have no conflicts of interest.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

1. Introduction

Expressed emotion (EE) is an interview-based measure of critical, hostile, emotionally overinvolved attitudes among caregiving family members toward a patient with a psychiatric disorder (Brown et al., 1972; Leff and Vaughn, 1984). Family members with elevations in these attitudes are given a designation of high (versus low) EE. Family EE is one of the most robust predictors of psychiatric relapse among patients with schizophrenia, bipolar disorder, and major depressive disorder (Butzlaff and Hooley, 1998; Hooley, 2007; Weintraub et al., 2017). High-EE attitudes can be broken down into subtypes, with caregivers identified as primarily critical/hostile, emotionally overinvolved (i.e., excessive self-sacrifice or overprotective), or both, potentially informing treatment of the ill family member. Warmth is an additional subcomponent of EE and, although receiving less attention in the research literature, has been found to be a protective factor in mitigating psychiatric relapse (López et al., 2004).

Despite its strong record in prospective studies, assessments of EE are rare in clinical settings. This is likely a result of the time-consuming and somewhat cumbersome nature of these assessments, making the process too time-intensive for clinicians in most settings. The gold-standard EE assessment – the semi-structured Camberwell Family Interview (CFI; Brown et al., 1972) – involves a 1–2 hour interview per family member and then an additional 3–4 hours to code for the dimensions of EE (Hooley and Parker, 2006). The rater codes the interview for critical, hostile, and/or overinvolved comments made by the speaker (usually a parent) regarding the subject (typically a patient/offspring). The coding requires expertise in the content required for high EE classification, careful consideration to the overall context of the comment, and acoustic tone and other paralinguistic features of speech. For example, the speaker is only deemed critical if their criticisms are directed at the subject, uses content such as “I don’t like it when…” and has a change in voice tone.

The Five-Minute Speech Sample (FMSS; Magaña et al., 1986) is an EE coding system that was developed to ease classification of EE relative to the CFI. The FMSS corresponds relatively well to CFI classifications for identifying high-EE attitudes (specificity=71%–91% & sensitivity=65.2%–80.0%) (Leeb et al., 1991; Magaña et al., 1986). While the FMSS improves measurement and coding burden relative to the CFI, it still requires that the rater attend a training workshop and undergo extensive reliability evaluations, and allot at least 20 minutes to code each sample (Hooley and Parker, 2006). Once the rater is deemed reliable, they must then allot at least 20 minutes to code each sample as well as conduct future reliability evaluations to ensure a maintenance of their reliability. Like the CFI, rating of the FMSS requires close attention to the content and the context of the speech. Thus, the FMSS remains too time-consuming and costly to provide at-scale in community mental health practices.

Some self-report measures, like the Perceived Criticism scale, have been developed to further ease assessment burden of EE. These measures provide the greatest advantage over Camberwell-based coding in regards to reducing the measurement and coding burden, although they appear to measure a slightly different construct. For example, the Perceived Criticism Scale, a commonly used 2-item scale, measures patient’s perception of their parent/caregiver's criticism as opposed to an objective third-party's determination of the caregiver's attitudes (Hooley and Teasdale, 1989; Masland et al., 2018).

Since EE is a reflection of individuals' speech, it is possible that EE could be measured by machine analysis of speech production. An automated method of categorizing EE and its subtypes through parents' speech samples could enable more widespread assessment of the construct in clinical settings. Additionally, the application of linguistic analyses to parental speech samples can help elucidate a potential "digital phenotype" of EE. Digital phenotyping aims to quantify behavioral phenotypes though features of voice and speech, smartphone sensors, or keyboard interactions (Insel, 2017). Language analysis tools, such as the Linguistic Inquiry and Word Count (LIWC) software (Pennebaker et al., 2007), extract information from textual data and sort text into various speech/language features, including linguistic properties (e.g., analytical thinking and emotional tone), word categories (e.g., pronouns, verbs, emotion words), and language markers (e.g., nonfluencies – "um," "er"). The LIWC cannot determine the context of the situation or acoustic tone of the speaker, but instead relies solely on linguistic properties (e.g., word categories) to code speech. LIWC coding has been used to detect individuals with suicidality, depression, neuroticism, and cognitive impairment from speech samples (Braithwaite et al., 2016; Jarrold et al., 2010; Resnik et al., 2013).

In this study, we evaluated whether parents' EE classification could be automated using machine learning of the lexical features of speech samples. We used FMSSs from parents of youth (ages 9–18) with mood disorders who enrolled in a randomized trial of family therapy, and examined the linguistic features that were associated with overall EE as well as the EE subtypes (criticism, emotional overinvolvement, and warmth) based on standard FMSS ratings from the Magana et al. (1986) system. Our primary goal was to derive a machine-learning algorithm that achieved predictive accuracy comparable to rates of inter-rater reliability for human coding of EE (kappa = 0.7–0.8; Leeb et al., 1991; Magaña et al., 1986). As an exploratory aim, we also examined which speech features related to overall EE and its subtypes in order to determine whether clarification of the EE construct could be garnered from machine learning analyses of speech.

2. Methods

2.1. Participants

A total of 126 parents of 92 youth provided speech samples for this study; 3 of the speech samples were inaudible and could not be rated on EE. Thus, the final sample included 123 parents with viable EE data. Participants for this study were parents of youth who were recruited for a multisite randomized trial of family-focused treatment (FFT) for youth at clinical and familial risk for bipolar disorder (BD). To be eligible for the trial, youth met the following criteria: (a) between ages 9 and 17 years, (b) had a DSM-IV-TR (American Psychiatric Association, 2000) diagnosis of BD not otherwise specified (NOS; currently termed other specified and related bipolar disorder in DSM-5; American Psychiatric Association, 2013) or major depressive disorder, as verified by Kiddie Schedule for Affective Disorders and Schizophrenia, Present and Lifetime version interviews and ratings (K-SADS; Axelson et al., 2003; Chambers et al., 1985; Kaufman et al., 1997); (c) had at least one first- or second-degree relative with a lifetime history of BD I or II as determined by Mini-International Neuropsychiatric Interview (MINI; Sheehan et al., 1998) or, when the relative was not available for direct interview, the Family History Screen (Weissman et al., 2000); and (d) had current elevations of mood symptoms (1-week Young Mania Rating Scale [YMRS] score>11 or 2-week Children's Depression Rating Scale, Revised [CDRS-R] score>29).

Clinical interviews were conducted by trained diagnosticians with at least a bachelor's degree. A study-affiliated psychiatrist conducted a separate diagnostic evaluation of the youth participant, and the youths' final primary diagnosis was determined based on consensus between the diagnostician and the study psychiatrist. The trial was conducted at the UCLA School of Medicine and the University of Colorado, Boulder, and approved by the human subjects review boards of both institutions. Each participant provided informed written consent or assent after receiving a complete description of the study.

2.2. Procedure

Upon meeting eligibility for the study, parents and/or stepparents of the children were asked to talk for five minutes each about "what kind of person is [*youth's name*], and how the two of you get along together," as outlined in the standardized FMSS task. The audio was recorded and an expert rater (Ana Magaña-Amato, MA) classified each parent as high or low EE, and identified the EE subtype (highly critical/hostile, emotionally overinvolved, both, or neither) based on the speech samples (Magaña et al., 1986). Parents were classified as high in overall EE for criticism if the speech sample contained an initial negative statement about the child, negative comments about the relationship with their child, or at least one critical comment based on the CFI criteria (negative content with change in acoustic tone or speed). Parents could also be classified as high EE if they evidenced emotional overinvolvement (EOI) by expressing an inordinate number of positive comments (i.e., 5 or more), had significant emotional reactions when describing the child's behavior, or described multiple instances of inordinate self-sacrifice (e.g., "I'd do anything in the world for my son"). Finally, regardless of their high- or low-EE status, parents are rated on a 'warmth' index based on expressions of positive regard, caring, concern or empathy. As per convention, families are classified as high-EE if one or more parents was high in critical comments or was classified as high in EOI (interrater reliabilities=0.82 and 0.80, respectively).

Each of the FMSS speech samples were transcribed and transcripts were analyzed using the Linguistic Inquiry Word Count (LIWC) software (available at www.liwc.net). The LIWC software generates a total of 93 speech features per sample including the word count of the transcript as well as the percentages of words that match different categories in its word library. For example, "anger" is a speech feature within the LIWC, which includes words like hate, kill, and annoyed. For each transcript, LIWC generates feature values (i.e., percentage of words that match a speech feature category) for each of the 93 speech features. Table 1 presents a sample FMSS transcript, and shows the LIWC process of classifying words into linguistic categories (in this case, positively and negatively-toned words) and calculating a numerical value for the proportion of words in each category.

2.3. Data Analysis

We conducted a total of four separate machine learning analyses to examine which speech features related to each of the EE classifications of parents: high vs. low EE, high vs. low criticism, high vs. low EOI, and high vs. low warmth, as described below. EE classifications were examined as dichotomous outcome variables. Parents who were missing a FMSS transcript were removed from the analyses.

We took multiple analytic steps to narrow down the 93 speech features that were most closely associated with EE (see Figure 1 for visual depiction of analytical process). First, we examined which linguistic features were univariately related to overall EE and EE subtypes. Twelve features that represent punctuation (e.g., apostrophes) were excluded. Then, we eliminated all speech features that did not meet one of the following criteria: (1) a significant (p<0.05 and Rho>0.1) pairwise Spearman rank correlation between speech feature and EE classification or (2) identification as stochastically relevant based on the Boruta algorithm. This algorithm uses decision trees to randomly sample speech features and iteratively determines which features are most important in predicting the dependent variable. Through this machine learning process, features that otherwise may be discarded by the pairwise comparisons are included in the predictive algorithms. The importance of each feature is measured at the end of each iteration and a ranking is created to select the most useful predictors (Kursa and Rudnicki, 2010). All of the above steps were implemented in R (Team, 2016). These initial univariate analyses led to a reduced set speech features for each of the EE outcomes (from the original 93 LIWC speech features) that were deemed preliminarily relevant for EE classification.

An updated dataset consisting of the EE ratings and the reduced speech features was then used as the input to a Support Vector Machine (SVM) model to find an optimal classifier for overall EE and each of the EE subtypes. SVMs have been shown to perform well as a classifier within similar studies of speech samples (Arevian et al., 2020). The SVMs for each EE classification were first tuned across the algorithm’s parameter space (cost function and sigma) using a radial kernel and 10-fold, resampled, Monte Carlo cross-validation. This method of validation involves repeated, random sub-sampling of the data. For each random sub-sampling of the data set, the model is fit to the data and predictive accuracy is assessed. The results are then averaged over each of the splits. The resulting SVM was used as the baseline for successive iterations where the reduced feature space was iteratively diminished in search of the maximal Receiver Operating Characteristic (ROC) curve and the minimal number of speech features. ROC-based evaluation uses class probabilities to find the optimal balance between sensitivity and specificity (Kuhn and Johnson, 2013), and can vary between 0.5 and 1, where 0.5 represents no classification is possible and 1 represents a perfect classification.

Once the maximum ROC was derived, the final SVM classification results were extracted, which included the features’ accuracy in predicting the EE classification, sensitivity and specificity of the prediction, the No Information Rate (i.e., accuracy achieved if classification was based on randomly predicting the most prevalent class), and the specific speech features in the classifier. All SVMs were implemented using the R-package kernlab (Karatzoglou et al., 2004).

Finally, in order to examine the validity of the machine learning classifications, we examined the relationship between the machine-derived classifications of overall EE and its subtypes with baseline youth manic and depressive symptoms on the K-SADS, parental distress on the Symptom Checklist-90-Revised (SCL-90; Derogatis, 1979), and family conflict on the parent-rated Conflict Behavioral Questionnaire (CBQ; Robin and Foster, 1995).

3. Results

3.1. Sample Characteristics

Of the 123 parents with viable EE data, 69 (56.1%) were rated as high EE based on their FMSS-rated classifications and 54 (43.9%) low-EE. A total of 82 (66.7%) of the caregivers were female, with 50 of these (61.0%) classified as high EE. The remaining 41 (33.3%) were male, with 19 of these (46.3%) classified as high EE ($X^2(1)=2.37$, $p=0.13$). There was no difference between male and female caregivers in being classified as high versus low EE status ($X^2(1)=1.34$, $p=0.25$).

The racial and ethnic make-up of the parents was not collected directly; however, the youth in the sample (*M* age=13.3 years; *SD*=2.6) were predominantly Caucasian (*n*=78; 84.8%) and 19 (20.6%) were Hispanic. The youth presented to the study with, on average, moderate-to-severe depressive symptom severity (*M* CDRS score=47.8; *SD*=15.6) and mild-to-moderate manic symptom severity (*M* YMRS score=13.5; *SD*=7.3). The majority of the youth (*n*=56; 60.9%) had a primary mood diagnosis of major depressive disorder; the remainder had other specified BD.

Of the 123 parents, 30 (24.4%) were rated as high in the criticism subtype of EE, 24 (19.5%) as high in the EOI subtype, and 15 (12.2%) as high in both. Additionally, 22 parents (17.9%) were rated high in warmth. Being rated high on emotional overinvolvement was not related to being high on criticism or warmth ($X^2(1)=0.18$, $p=0.69$; $X^2(1)=2.23$, $p=0.14$, respectively). However, parents who were rated as highly critical were less likely to be rated as high in warmth ($X^2(1)=11.37$, $p<0.001$).

3.2. Predictive accuracy of LIWC features on overall EE ratings

The 123 parental speech samples contained a mean of 718.5 words (*SD*=158.8), with no sex differences among parents. In the initial feature identification phase for overall EE (using the Spearman rank correlations and Boruta algorithm), a total of 16 LIWC speech features were selected as being individually predictive of overall (high versus low) EE (see Table 2). Use of informal language (a category of words that include swear words, netspeak, and nonfluencies) as well as the specific nonfluencies speech feature (e.g., “er” and “um”) had the strongest relationship with EE, both associated negatively with EE. Anger words (e.g., “hate,” “annoyed”) and references to male words (e.g., “he”, “boy”) were strongly positively related to high EE status.

The final step for EE classification used support vector machine (SVM) classification based on the initially identified 16 speech features. The SVM used a total of 10 of the previously identified features to achieve the highest degree of predictive accuracy: 75.2% accuracy in predicting high versus low EE, with a sensitivity of 0.69 and specificity of 0.81. This accuracy rate is higher than the No Information Rate of 56% (i.e., the base rate of the most frequently occurring category – which in this case is high EE).

3.3. Predictive accuracy of LIWC features on criticism

The initial LIWC feature selection analysis for the criticism subtype of EE identified 15 features as individually relevant predictors (see Table 3). The speech features that most strongly related to criticism were anger words, negative emotion words (a word category that includes anger, anxiety, and sadness words), and tentative words (e.g., “perhaps,” “maybe”).

The SVM classification used 6 of the previously identified 15 features to achieve its optimal prediction of criticism. The SVM achieved 75.5% accuracy, with a sensitivity of 0.63 and a specificity of 0.83. This prediction rate is higher than the No Information Rate of 63%. A comparison of features that were associated with overall EE and each EE subtype is presented in Table 4.

3.4. Predictive accuracy of LIWC features on emotional overinvolvement (EOI)

In the initial classification of the EOI subtype of EE, a total of 14 LIWC speech features were selected as individually predictive of EOI (see Table 3). The speech features of perceptual processes (e.g., “look,” “heard”) and see (e.g., “view,” “saw”) had the strongest associations with high EOI, both being positively associate with EOI. Auxiliary verbs (e.g., “am,” “will”) and biological processes words (e.g., “eat,” “pain”) also had strong positive associations with EOI, whereas greater use of informal language and nonfluencies were negatively associated with EOI.

The final SVM classification used all 14 of the previously identified speech features to most accurately predict EOI. This SVM achieved 77.3% accuracy, with a sensitivity of 0.67 and a specificity of 0.82. This prediction rate is higher than the No Information Rate of 68%, the base rate of low EOI in this sample.

3.5. Predictive accuracy of LIWC features on warmth

A total of 10 LIWC speech features were initially identified as predictive of the warmth subtype of EE (see Table 3). Negative emotion words and negate words (“no,” “never”) had the strongest negative association with warmth.

The final SVM classification used 5 of the previously identified features and achieved 81.8% accuracy, with a sensitivity of 0.74 and a specificity of 0.89. This prediction rate is equal to the No Information Rate of 82%.

3.6. Machine-derived EE classifications in relation to youth clinical and family features

The machine-derived parental criticism subtype was related to greater concurrent manic symptoms ($F(1,109)=6.00$, $p=0.02$) as well as greater parent-rated distress and family conflict ($F(1,80)=8.20$, $p=0.01$; $F(1,117)=6.05$, $p=0.02$, respectively). In parallel, observer ratings parental criticism using the FMSS were related to more severe manic symptoms, parental distress, and family conflict ($F(1,109)=4.36$, $p=0.04$; $F(1,80)=4.12$, $p=0.05$;

F(1,117)=7.87, *p*=0.01, respectively). Emotional overinvolvement and overall classifications of EE (high vs. low) from both the observer ratings and the machine-learning process did not relate to youth mood, parental distress or family conflict.

4. Discussion

We examined the accuracy of machine learning algorithms in classifying the EE status of parents of youth with mood disorders. We first identified speech features parents' Five-Minute Speech Samples (FMSS) that had significant pair-wise relationships with their EE classifications. We then used support vector machine (SVM) algorithms to examine the most accurate combination of these speech features in predicting EE. The final algorithms produced high accuracy levels in classifying overall (high versus low) EE and EE subtypes (≥0.75). While both sensitivity and specificity values indicated high predictive rates (>0.6 and >0.8, respectively), the models favored specificity over sensitivity (i.e., reducing the misclassifications of parents as high-EE who were actually low EE). Additionally, there was evidence for clinical validity of the machine-learning classifications, as the machine-derived criticism subtype related to other clinical and family functioning measures. Together, machine learning analysis of lexical speech features shows promise as an efficient method of classifying parents into high and low-EE categories.

The machine learning algorithms achieved predictive accuracy levels that were comparable to interrater reliabilities (kappas=0.7–0.8) for EE (Magaña et al., 1986). This study represents a first step toward developing an automated method of EE classification in community practice. Once a replicable, evidence-based algorithm is created for EE classification, the process of determining an individual's EE status could be automated through software that combines audio-to-text transcription of the FMSS, extraction of lexical features, and running of the SVM classifiers. Once this process is automated, coding EE could take as little time per family as it takes to record the FMSS, eliminating the need for specialized training of staff and time to manually code the transcripts.

Assessing the EE status of a family of a teen with psychiatric disorder has implications for family interventions such as family-focused therapy (FFT) (Miklowitz and Chung, 2016). FFT seeks to reduce conflict and improve cohesion among family members through use of modules pertaining to psychoeducation, communication enhancement training, and problem-solving. However, not all families require the same level of training in communication and problem-solving skills. Several studies indicate that the effects of FFT on symptom outcomes are greater in patients with or at risk for BD who reside in high-EE families compared to those in low-EE families (Kim and Miklowitz, 2004; Miklowitz et al., 2009; Miklowitz et al., 2013). Families with low levels of EE may benefit from truncated versions of FFT psychoeducation (Miklowitz et al., 2014; Miklowitz et al., 2020), whereas those with high criticism may benefit from communication training that focuses on increasing positive communication and conflict resolution. The needs of families with high overinvolvement include a greater focus on assisting parents in developing broader social support networks or encouraging greater independence of the patient.

The results of this study help identify speech features that are and are not relevant to overall EE and EE subtypes. As expected, negative emotion words (e.g., “hate”) were positively associated with high EE and negatively associated with warmth. This finding may be related to high EE parents’ greater self-reported anger/hostility as compared to low EE parents, which represents a critical target for intervention (Millman et al., 2018). Despite EOI appearing to be an anxious response to the patient’s illness and/or concerns about relapse, parental EOI was not reflected in words pertaining to anxiety (e.g., “nervous”). Instead, the presence of EOI was connected to positive emotion words such as “love” or “nice.” The EOI construct has multiple dimensions, some of which may be protective for the patient (e.g., appropriate levels of self-sacrifice among parents) and some may contribute to risk (inappropriate emotional responses to symptoms) (Fredman et al., 2015). In addition to the speech features capturing the parents’ own emotions, the parents are commonly describing their youth’s emotions, cognitions, and behaviors in the FMSS (e.g., “my son is so anxious”). Thus, the identified speech features appear to also represent parents’ perceptions of affective, behavioral, or cognitive processes in the youth.

A speech feature that has been previously identified in the speech literature and, in this study, distinguished the critical and EOI subtypes is tentative speech. Tentative speech contains pauses in speech or qualifiers (e.g., maybe), which may convey attempts to be polite or to acknowledge inadequate information on which to base a judgement (Gibbons et al., 1991; Holtgraves and Lasky, 1999). Within the parents’ FMSSs, high-EE, tentative speech can serve to reduce the appearance of anger or hostility towards the offspring during an interview with a clinical researcher. Alternatively, some parents may be uncertain how to feel when balancing their frustration with the knowledge that their child has a psychiatric disorder.

4.1. Limitations

Analyses of speech transcripts are limited via LIWC because they do not consider the acoustic tone of the speaker as well as the ability to comprehend the context of the conversation. In contrast, the human coding of EE from the Camberwell Family Interview (CFI) or the FMSS requires careful consideration of changes in voice tone when discussing the patient’s negative or provocative behaviors. Analyses of speech that considers paralinguistic information may improve the accuracy of these algorithms in relation to rater-based coding. This study is also limited by the reliance on the FMSS coding system as the “ground truth” by which machine learning produced classification algorithms. The CFI is considered the gold standard measurement of EE, with the FMSS under-identifying high EE relatives (Hooley and Parker, 2006). Thus, automated algorithms for classifying EE and EE subtypes may have less sensitivity when compared to CFI-based ratings.

Our results require replication to ensure that the same algorithms produce high predictive accuracy in independent samples. Machine learning analyses that seek to optimize parameters and variable selections (as was the goal here) would ideally perform analyses in three parts – training, validation, and test. The training step uses machine learning to derive an algorithm on an initial sample and then the algorithm’s accuracy sensitivity, and specificity are validated in that same sample. These first two steps were done in this paper.

An important third step would be to test the algorithm on an withheld portion of the sample (or an independent sample) to test whether sensitivity and specificity are consistent in a new sample. This was not possible in this study due to limitations of sample size. A potential consequence of omitting the test-set step is overfitting (Vabalas et al., 2019). As a result, the degree to which the machine learning's algorithm overfits this sample is unknown. Additionally, there could be a feature selection bias due to selecting speech features before cross-validation, rather than nesting the feature selection process within the cross-validation. To reduce potential bias in the feature selection process, we conducted a Boruta analysis, which is faster than the nested feature selection implementation and provides an unbiased selection (Kursa and Rudnicki, 2010). We felt that this was a good compromise given the exploratory nature of this study, but future research should examine the reliability of our findings through additional speech feature selections processes.

4.2. Future Directions

This sample focused on parents of youth with mood disorders. It is unclear whether these algorithms will replicate in caregivers of patients who are older or among patients with other psychiatric disorders (e.g., schizophrenia). There may also be aspects of EE expression that are unique to parents (notably, emotional overinvolvement) as opposed to spouses. Future research should examine whether the same machine learning algorithms can be applied to classify EE across illness, patient populations, and types of family members/caregivers. Research in psychiatry is beginning to leverage machine learning of speech samples to track patients' clinical states (Arevian et al., 2020), and is moving towards using these technologies to determine patients' diagnoses and clinical prognosis (Insel, 2017). Due to the impact of family environment on mental health (Hooley, 2007), it will be important to examine the validity of machine-derived EE classifications as predictors of traditional clinical outcomes, such as illness recurrences. It will also be useful to add clarity to the EE construct by examining why certain speech features are correlated with high versus low EE ratings (e.g., space words, ingest words). It is unclear whether these features might have direct clinical meaning, are indirectly related to clinically meaningful constructs (e.g., personality attributes of the speaker) or are just an artifact of this sample.

5. Conclusions

Study findings suggest that machine learning processes can be successfully applied to classify parents' overall EE status and their EE subtypes. Once replicated, EE classification via parent speech samples can then be automated to promote more efficient assessment of the family climate. Additionally, the speech features that were elucidated represent an initial step in creating a digital phenotype for EE classifications. Whereas some of the speech features that were associated with EE classifications are consistent with previous research (e.g., negative emotions, anger), many of the speech features did not have a clear, a priori connection to EE (e.g., nonfluencies). The tool of machine learning can support the process of automating of EE measurement as well as highlight speech features for future investigation.

Role of funding sources:

Financial support for this study was provided by National Institute of Mental Health (NIMH) grants R01MH093676, R01MH093666, R34MH077856, and R34MH117200, and AIM For Mental Health.

References

American Psychiatric Association, 2000. DSM-IV-TR: Diagnostic and statistical manual of mental disorders, text revision. Washington, DC: American Psychiatric Association.

American Psychiatric Association, 2013. Diagnostic and statistical manual of mental disorders (DSM-5®). American Psychiatric Pub.

Arevian AC, Bone D, Malandrakis N, Martinez VR, Wells KB, Miklowitz DJ, Narayanan S, 2020. Clinical state tracking in serious mental illness through computational analysis of speech. PLoS one 15(1), e0225695. [PubMed: 31940347]

Axelson D, Birmaher BJ, Brent D, Wassick S, Hoover C, Bridge J, Ryan N, 2003. A preliminary study of the Kiddie Schedule for Affective Disorders and Schizophrenia for School-Age Children mania rating scale for children and adolescents. Mary Ann Liebert, Inc.

Braithwaite SR, Giraud-Carrier C, West J, Barnes MD, Hanson CL, 2016. Validating machine learning algorithms for Twitter data against established measures of suicidality. JMIR mental health 3(2), e21. [PubMed: 27185366]

Brown GW, Birley J, Wing JK, 1972. Influence of family life on the course of schizophrenic disorders: a replication. The British Journal of Psychiatry.

Butzlaff RL, Hooley JM, 1998. Expressed emotion and psychiatric relapse: a meta-analysis. Archives of general psychiatry 55(6), 547–552. [PubMed: 9633674]

Chambers WJ, Puig-Antich J, Hirsch M, Paez P, Ambrosini PJ, Tabrizi MA, Davies M, 1985. The assessment of affective disorders in children and adolescents by semistructured interview: test-retest reliability of the Schedule for Affective Disorders and Schizophrenia for School-Age Children, Present Episode Version. Archives of general psychiatry 42(7), 696–702. [PubMed: 4015311]

Derogatis LR, 1979. Symptom Checklist-90-Revised (SCL-90-R). Lyndhurst, NJ: NCS Pearson.

Fredman SJ, Baucom DH, Boeding SE, Miklowitz DJ, 2015. Relatives’ emotional involvement moderates the effects of family therapy for bipolar disorder. Journal of consulting and clinical psychology 83(1), 81. [PubMed: 25198285]

Gibbons P, Busch J, Bradac JJ, 1991. Powerful versus powerless language: Consequences for persuasion, impression formation, and cognitive response. Journal of Language and Social Psychology 10(2), 115–133.

Holtgraves T, Lasky B, 1999. Linguistic power and persuasion. Journal of Language and Social Psychology 18(2), 196–205.

Hooley JM, 2007. Expressed emotion and relapse of psychopathology. Annu. Rev. Clin. Psychol. 3, 329–352. [PubMed: 17716059]

Hooley JM, Parker HA, 2006. Measuring expressed emotion: An evaluation of the shortcuts. Journal of Family Psychology 20(3), 386. [PubMed: 16937995]

Hooley JM, Teasdale JD, 1989. Predictors of relapse in unipolar depressives: Expressed emotion, marital distress, and perceived criticism. Journal of abnormal psychology 98(3), 229. [PubMed: 2768657]

Insel TR, 2017. Digital phenotyping: technology for a new science of behavior. Jama 318(13), 1215–1216. [PubMed: 28973224]

Jarrold WL, Peintner B, Yeh E, Krasnow R, Javitz HS, Swan GE, 2010. Language analytics for assessing brain health: Cognitive impairment, depression and pre-symptomatic Alzheimer’s disease, International Conference on Brain Informatics. Springer, pp. 299–307.

Karatzoglou A, Smola A, Hornik K, Zeileis A, 2004. kernlab-an S4 package for kernel methods in R. Journal of statistical software 11(9), 1–20.

Kaufman J, Birmaher B, Brent D, Rao U, Flynn C, Moreci P, Williamson D, Ryan N, 1997. Schedule for affective disorders and schizophrenia for school-age children-present and lifetime version (K-

SADS-PL): initial reliability and validity data. Journal of the American Academy of Child & Adolescent Psychiatry 36(7), 980–988. [PubMed: 9204677]

Kim EY, Miklowitz DJ, 2004. Expressed emotion as a predictor of outcome among bipolar patients undergoing family therapy. Journal of Affective Disorders 82(3), 343–352. [PubMed: 15555685]

Kuhn M, Johnson K, 2013. Applied predictive modeling. Springer.

Kursa MB, Rudnicki WR, 2010. Feature selection with the Boruta package. J Stat Softw 36(11), 1–13.

Leeb B, Hahlweg K, Goldstein MJ, Feinstein E, Mueller U, Dose M, Magana-Amato A, 1991. Cross-national reliability, concurrent validity, and stability of a brief method for assessing expressed emotion. Psychiatry Research 39(1), 25–31. [PubMed: 1771207]

Leff J, Vaughn C, 1984. Expressed emotion in families: Its significance for mental illness. Guilford Press.

López SR, Nelson Hipke K, Polo AJ, Jenkins JH, Karno M, Vaughn C, Snyder KS, 2004. Ethnicity, expressed emotion, attributions, and course of schizophrenia: family warmth matters. Journal of Abnormal Psychology 113(3), 428. [PubMed: 15311988]

Magaña AB, Goldstein MJ, Karno M, Miklowitz DJ, Jenkins J, Falloon IR, 1986. A brief method for assessing expressed emotion in relatives of psychiatric patients. Psychiatry research 17(3), 203–212. [PubMed: 3704028]

Masland SR, Drabu S, Hooley JM, 2018. Is perceived criticism an independent construct? Evidence for divergent validity across two samples. Journal of Family Psychology.

Miklowitz DJ, Axelson DA, George EL, Taylor DO, Schneck CD, Sullivan AE, Dickinson LM, Birmaher B, 2009. Expressed emotion moderates the effects of family-focused treatment for bipolar adolescents. Journal of the American Academy of Child & Adolescent Psychiatry 48(6), 643–651. [PubMed: 19454920]

Miklowitz DJ, Chung B, 2016. Family-focused therapy for bipolar disorder: Reflections on 30 years of research. Family process 55(3), 483–499. [PubMed: 27471058]

Miklowitz DJ, Schneck CD, George EL, Taylor DO, Sugar CA, Birmaher B, Kowatch RA, DelBello MP, Axelson DA, 2014. Pharmacotherapy and family-focused treatment for adolescents with bipolar I and II disorders: a 2-year randomized trial. American Journal of Psychiatry 171(6), 658–667.

Miklowitz DJ, Schneck CD, Singh MK, Taylor DO, George EL, Cosgrove VE, Howe ME, Dickinson LM, Garber J, Chang KD, 2013. Early intervention for symptomatic youth at risk for bipolar disorder: a randomized trial of family-focused therapy. Journal of the American Academy of Child & Adolescent Psychiatry 52(2), 121–131. [PubMed: 23357439]

Miklowitz DJ, Schneck CD, Walshaw PD, Singh MK, Sullivan AR, Suddath RL, Forgey-Borlik M, Sugar CA, Chang K, 2020. Family-focused Therapy for Symptomatic Youths at High Risk for Bipolar Disorder A Randomized Clinical Trial. JAMA Psychiatry.

Millman ZB, Weintraub MJ, Miklowitz DJ, 2018. Expressed emotion, emotional distress, and individual and familial history of affective disorder among parents of adolescents with bipolar disorder. Psychiatry research 270, 656–660. [PubMed: 30384286]

Pennebaker JW, Booth RJ, Francis ME, 2007. Linguistic inquiry and word count: LIWC [Computer software]. Austin, TX: liwc. net 135.

Resnik P, Garron A, Resnik R, 2013. Using topic modeling to improve prediction of neuroticism and depression in college students, Proceedings of the 2013 conference on empirical methods in natural language processing. pp. 1348–1353.

Robin A, Foster S, 1995. The conflict behavior questionnaire. Dictionary of Behaviorial Assessment Techniques. New York: Pergamon, 148–150.

Sheehan DV, Lecrubier Y, Sheehan KH, Amorim P, Janavs J, Weiller E, Hergueta T, Baker R, Dunbar GC, 1998. The Mini-International Neuropsychiatric Interview (MINI): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. The Journal of clinical psychiatry.

Team, R.C., 2016. R: A language and environment for statistical computing.

Vabalas A, Gowen E, Poliakoff E, Casson AJ, 2019. Machine learning algorithm validation with a limited sample size. PloS one 14(11), e0224365. [PubMed: 31697686]

Weintraub MJ, Hall DL, Carbonella JY, Weisman de Mamani A, Hooley JM, 2017. Integrity of literature on expressed emotion and relapse in patients with schizophrenia verified by a p-curve analysis. Family process 56(2), 436–444. [PubMed: 26875506]

Weissman MM, Wickramaratne P, Adams P, Wolk S, Verdeli H, Olfson M, 2000. Brief screening for family psychiatric history: the family history screen. Archives of General Psychiatry 57(7), 675–682. [PubMed: 10891038]

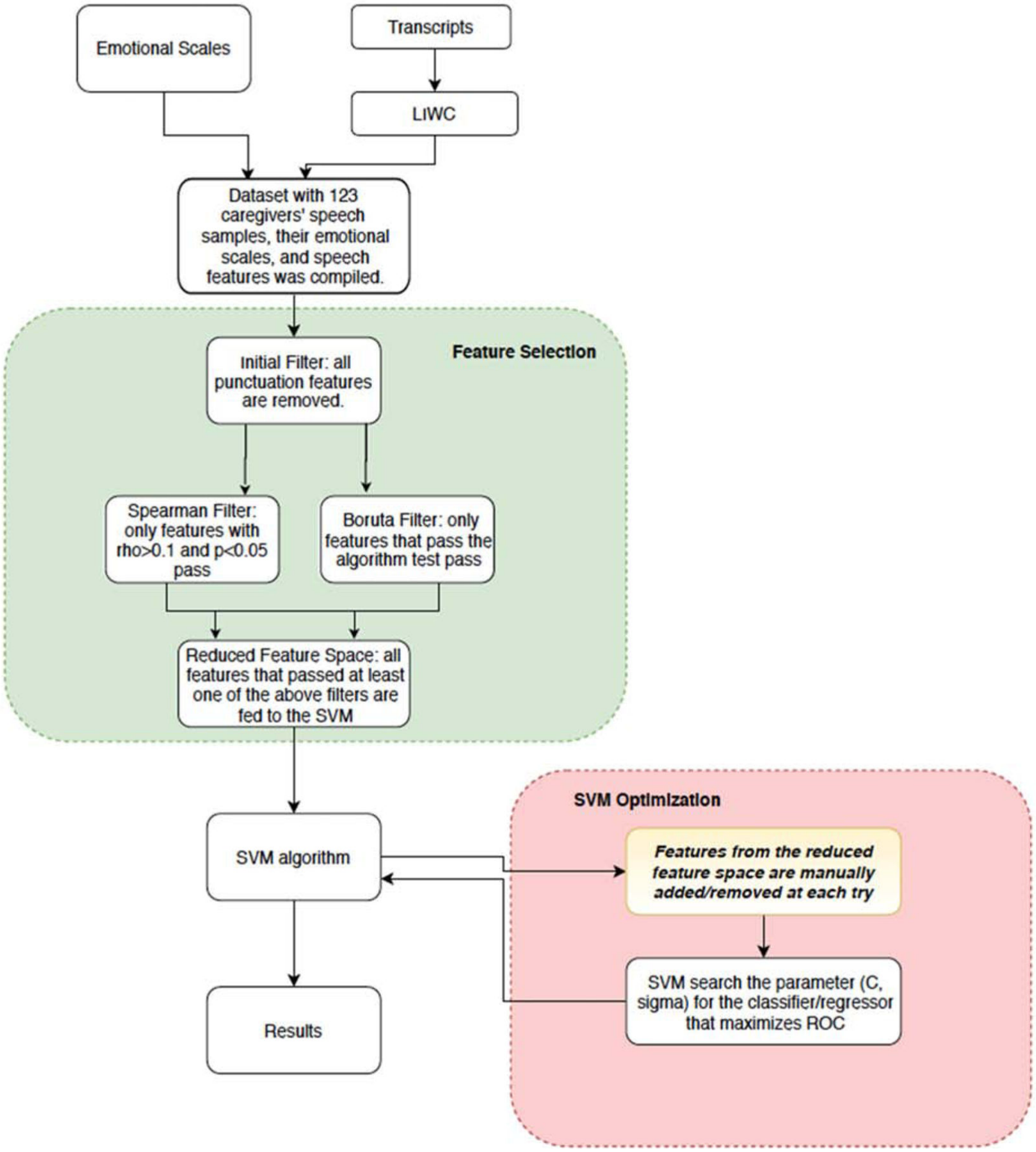
Emotional Scales
Transcripts
LIWC
Dataset with 123 caregivers' speech samples, their emotional scales, and speech features was compiled.
Feature Selection
Initial Filter: all punctuation features are removed.
Spearman Filter: only features with rho>0.1 and p<0.05 pass
Boruta Filter: only features that pass the algorithm test pass
Reduced Feature Space: all features that passed at least one of the above filters are fed to the SVM
SVM Optimization
SVM algorithm
Features from the reduced feature space are manually added/removed at each try
SVM search the parameter (C, sigma) for the classifier/regressor that maximizes ROC
Results

Figure 1.

Analytical process for each of the four EE classification analyses

Table 1.

Sample Transcript and Associated LIWC Speech Features

"….Um, but I guess in terms of our relationship it's pretty, yeah and I admit that I'm not always um, appropriate in terms of my interaction with her. I get very irritated and I disengage and then that makes her anxious and she comes back and it keeps, we keep going at it. So, um, I don't know. I think the biggest thing for me is that I, um, like I said I never wanted to acknowledge that, it was easier for me to say that she was anxious than to say that she has a mood disorder. And I feel guilty about that. And I see the symptoms. I'm in the mental health field and I know what to look for. And I don't want her to live another 10, 20 years being just miserable because I don't think she's happy. So I feel really guilty about that. I think that's why sometimes I push her away and I get irritable because I want to think it's something different. I want to think it's behavioral. I want to think that it's just normal teen stuff and I know it's not……"

Word Count (WC)	Positive Words	Negative Words	Posemo	Negemo	Affect
201	3	7	1.49	3.48	5.97

The table summarizes the non-copyrighted LIWC results involving positive emotion, negative emotion words, and word count. The green highlighted words represent positive emotion words identified by the LIWC. The red highlighted words represent the negative emotions words identified in LIWC. Posemo and Negemo are the percentages of positive and negative emotion words (respectively) in the transcript. Affect is the sum of those two percentages.

Table 2.

LIWC speech features associated with overall Expressed Emotion (EE) classification based on a significant Spearman rank correlation or identification from the Boruta algorithm

	Spearman rank correlation	Cohen's d (95% CI)	*p*-value
Affect words *	0.21	−0.34 (−0.71, 0.02)	0.02
Anger *	0.22	−0.51 (−0.87, −0.14)	0.01
Authentic	−0.18	0.47 (0.11, 0.84)	0.04
Function words *	0.14	−0.31 (−0.67, 0.06)	0.13
Home	0.23	−0.47 (−0.84, −0.11)	0.01
Informal language *	−0.29	0.62 (0.25, 0.99)	0.001
Ingest *	0.20	−0.30 (−0.66, 0.06)	0.03
Insight *	−0.22	0.49 (0.13, 0.86)	0.02
Male references	0.27	−0.47 (−0.83, −0.10)	0.002
Negative emotion *	0.21	−0.31 (−0.67, 0.05)	0.02
Nonfluencies *	−0.28	0.59 (0.22, 0.96)	0.001
Relativity *	0.03	−0.09 (−0.45, 0.27)	0.72
See *	0.19	−0.39 (−0.75, −0.02)	0.04
Social *	0.20	−0.45 (−0.81, −0.08)	0.03
Space *	0.21	−0.43 (−0.79, −0.07)	0.02
Third-person singular	0.22	−0.46 (−0.83, −0.10)	0.02

* indicates speech features used in final SVM classification for EE

Table 3.

LIWC speech features associated with EE subtypes based on a significant Spearman rank correlation or identification from the Boruta algorithm

	Spearman rank correlation	Cohen’s d (95% CI)	*p*-value
Criticism			
Adjectives	−0.20	0.41 (0.03, 0.78)	0.02
Adverb	−0.18	0.26 (−0.12, 0.63)	0.04
Anger*	0.41	−0.90 (−1.29, −0.51)	<0.001
Anxiety*	0.05	0.01 (−0.36, 0.38)	0.55
Differentiation	0.18	−0.37 (−0.74, 0.00)	0.04
Female references	−0.18	0.39 (0.01, 0.76)	0.04
Informal language*	−0.20	0.41 (0.03, 0.78)	0.03
Insight*	−0.17	0.41 (0.04, 0.79)	0.06
Male references	0.20	−0.47 (−0.85, −0.10)	0.03
Negate*	0.20	−0.32 (−0.70, 0.05)	0.02
Negative emotions	0.25	−0.41 (−0.78, −0.03)	0.005
Nonfluencies	−0.19	0.34 (−0.03, 0.71)	0.04
Risk	0.18	−0.35 (−0.72, 0.03)	0.05
Tentative*	0.28	−0.52 (−0.90, −0.15)	0.002
Tone (emotional tone)	−0.19	0.36 (−0.01, 0.74)	0.04
Emotional Overinvolvement			
Affect words*	0.22	−0.44 (−0.83, −0.05)	0.01
Auxiliary verbs*	0.26	−0.60 (−0.99, −0.21)	0.004
Biological processes*	0.24	−0.64 (−1.03, −0.24)	0.007
Function words*	0.19	−0.41 (−0.80, −0.03)	0.03
Informal language*	−0.27	0.61 (0.22, 1.00)	0.003
Male references*	0.19	−0.32 (−0.71, 0.07)	0.04
Nonfluencies*	−0.24	0.62 (0.23, 1.01)	0.007
Perceptual processes*	0.31	−0.68 (−1.08, −0.29)	<0.001
Personal pronouns*	0.23	−0.51 (−0.90, −0.12)	0.01
Positive emotions*	0.18	0.52 (0.13, 0.91)	0.05
See*	0.28	−0.67 (−1.06, −0.28)	0.002
Six-letter words*	−0.17	0.43 (0.05, 0.82)	0.07
Tentative*	−0.22	0.52 (0.13, 0.91)	0.02
Verbs*	0.19	−0.45 (−0.84, −0.06)	0.04
Warmth			
Affiliation	0.18	−0.33 (−0.80, 0.14)	0.04
Anxiety*	−0.23	0.50 (−0.07, 0.87)	0.01

	Spearman rank correlation	Cohen’s d (95% CI)	*p*-value
Body	−0.22	0.54 (0.07, 1.01)	0.01
Filler	0.14	−0.36 (−0.83, 0.11)	0.12
Negate *	−0.25	0.63 (0.15, 1.10)	0.007
Negative emotions *	−0.28	0.75 (0.28, 1.23)	0.002
Nonfluencies *	0.02	−0.10 (−0.57, 0.36)	0.81
Positive emotions *	0.10	−0.18 (−0.65, −0.28)	0.28
Tone	0.22	−0.56 (−1.04, −0.09)	0.02
Word count	0.20	−0.47 (−0.94, −0.00)	0.03

* indicates speech features used in final SVM classification for that EE subtype

Table 4.

Associations of LIWC speech features with EE and EE subtypes

		EE	Criticism	EOI	Warmth
Speech features	*Definition or example words*				
Adjectives	Descriptor of nouns		−0.20		
Adverb	Descriptor of verbs		−0.18		
Affect words	Category of words that include positive and negative emotions	**0.21**		**0.22**	
Affiliation	“ally,” “friend,” “social”				0.18
Anger	“hate,” “kill,” “annoyed”	**0.22**	**0.41**		
Anxiety	“worried,” “fearful”		**0.05**		**−0.23**
Authentic	LIWC summary feature	−0.18			
Auxiliary verbs	“am,” “will,” “have”			**0.26**	
Biological processes	“eat,” “blood,” “pain”			**0.24**	
Body	“cheek,” “hands,” “spit”				−0.22
Differentiation	“hasn’t,” “but,” “else”		0.18		
Female references	“girl”, “her”, “mom”		−0.18		
Filler	“I mean,” “you know”				0.14
Function words	Word category that includes pronouns	**0.14**		**0.19**	
Home	“kitchen,” “landlord”	0.23			
Informal language	Word category that includes swear words, netspeak (e.g., lol), assent, nonfluencies, and fillers	**−0.29**	**−0.20**	**−0.27**	
Ingest	“dish,” “eat,” “pizza”	**0.20**			
Insight	“think,” “know”	**−0.22**	−0.17		
Male references	“boy,” “his,” “dad”	0.27	0.20	**0.19**	
Negate	“no,” “not,” “never”		**0.20**		**−0.25**
Negative emotion	Word category that includes anxiety, anger, and sadness words	**0.21**	0.25		**−0.27**
Nonfluencies	“er,” “hm,” “um”	**−0.28**	−0.19	**−0.24**	**0.02**
Perceptual processes	“look,” “heard,” “feeling”			**0.31**	
Personal pronouns	“I”, “we,” you,” “she/he,” “they”			**0.23**	
Positive emotion	“love,” “nice,” “sweet”			**0.18**	**0.10**
Relativity	“area”, “bend,” “exit”	**0.03**			
Risk	“danger,” “doubt”		0.18		
See	“view,” “saw,” “seen”	**0.19**		**0.28**	
Six-letter words	Words with 6 (or more) letters			**−0.17**	
Social processes	Word category that includes family, friends, and male/female referent words	**0.20**			
Space	“down,” “in,” “thin”	**0.21**			
Tentative	“maybe,” “perhaps”		**0.28**	**−0.22**	
Third-person singular	“she,” “he”	0.22			
Tone (emotional)	LIWC summary feature		−0.19		0.22
Verbs	Word that describes action or occurrence			**0.19**	
Word count	Number of words in speech sample				0.20

Values represent Spearman correlations between EE (or EE subtype) and LIWC speech features. All values that are greater than the absolute value of 0.17 are significant at $p<0.05$. Bolded features represent those used in final SVM classification.