Comparing emotion feature extraction approaches for predicting depression and anxiety

Hannah A. Burkhardt
University of Washington
haalbu@uw.edu

Michael D. Pullmann
University of Washington
pullmann@uw.edu

Thomas D. Hull
Talkspace
derrick@talkspace.com

Patricia A. Areán
University of Washington
parean@uw.edu

Trevor Cohen
University of Washington
cohenta@uw.edu

Abstract

The increasing adoption of message-based behavioral therapy enables new approaches to assessing mental health using linguistic analysis of patient-generated text. Word counting approaches have demonstrated utility for linguistic feature extraction, but deep learning methods hold additional promise given recent advances in this area. We evaluated the utility of emotion features extracted using a BERT-based model in comparison to emotions extracted using word counts as predictors of symptom severity in a large set of messages from text-based therapy sessions involving over 6,500 unique patients, accompanied by data from repeatedly administered symptom scale measurements. BERT-based emotion features explained more variance in regression models of symptom severity, and improved predictive modeling of scale-derived diagnostic categories. However, LIWC categories that are not directly related to emotions provided valuable and complementary information for modeling of symptom severity, indicating a role for both approaches in inferring the mental states underlying patient-generated language.

1 Introduction

Almost 10% of adults in the United States receive mental health counseling (Zablotsky and Terlizzi, 2020). The principle of measurement-based care dictates that medical treatments should be initiated and evaluated over time based on repeated assessments of patient symptoms and symptom trajectory (Scott and Lewis, 2015). In the context of talk therapy, mental health practitioners estimate treatment progress based on patients' current and historical verbal communications. For evaluating depression and anxiety severity, expressions of emotional state are key aspects of such communications (Beck, 1967; Rottenberg, 2017; Amstadter, 2008).

While prior work predominantly focused on sentiment, i.e. positive/negative polarity, expression of fine-grained emotions (Chancellor and De Choudhury, 2020; Guntuku et al., 2017) may give further insights into depression and anxiety symptomatology. For example, pride may be impacted by depression in a unique way. Gruber et al. (2011) showed that pride, a positive emotion relating to the self, is inversely correlated with depression, which is often associated with a poor self-image. At the same time, they found a smaller effect on joy and amusement, concluding that grouping these emotions into "positive affect" may result in a loss of nuance.

The increasing adoption of digital mental health tools and services, particularly message-based therapy, has afforded new opportunities to assist practitioners in quantifying depression and anxiety severity by assessing emotion in patient-generated text. Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2007; Tausczik and Pennebaker, 2010) is a software package designed to count words belonging to pre-defined categories with an extensive track record of validation for the detection of linguistic indicators of mental state (Tausczik and Pennebaker, 2010). It is commonly used to measure positive and negative affect, a limited set of specific emotions (sadness, anxiety, and anger), and other linguistic dimensions related to style and topic. Several LIWC categories have established relationships with depression, including the affect category sadness (e.g. "sad", "cry", "suffer"), the topic category health (e.g. "alcohol", "rash", "self-care"), and the syntactic category first-person pronouns (e.g. "I", "me", "my"). LIWC has been used to measure depression levels in social media posts (Coppersmith et al., 2014; De Choudhury et al., 2014, 2013a,b), therapy conversations (Burkhardt et al., 2021; Sonnenschein et al., 2018), and other written texts (Rude et al., 2004; Wiltsey Stirman and Pennebaker, 2001). LIWC measurements have also been shown to distinguish between patients with depression and those with anxiety

Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology, pages 105 - 115
July 15, 2022

disorders (Sonnenschein et al., 2018), correlate with self-reported measures of anxiety and worry in written descriptions of emotional responses to COVID-19 (Kleinberg et al., 2020), and predict whether posts emanated from anxiety-related subreddits (Shen and Rudzicz, 2017).

However, word counting methods cannot address linguistic phenomena such as negation ("not bad"), sarcasm, and context-dependence (for example, in the case of polysemy, words have multiple meanings that can only be disambiguated in context), and manually defined dictionaries may omit synonyms for terms they encode. Prior work suggests that neural network (NN)-based natural language processing (NLP) techniques can account for such phenomena and may therefore improve upon this straightforward word-counting method in their ability to identify concepts related to symptom severity. Shen and Rudzicz found that the performance of machine learning models identifying whether or not Reddit posts were drawn from anxiety-related subreddits improved when these models included neural word embeddings rather than LIWC-derived features (2017). However, the distributed representations of posts used in this work do not relate directly to interpretable emotion features. Further, contemporary transformer-based NN language models offer advantages over neural word embeddings in their ability to leverage proximal cues (such as "not") when interpreting the contextual meaning of a word. As noted by the authors, this work suggests a need for further research on automated assessments of linguistic indicators of anxiety disorders, involving larger data sets and explicit diagnostic assessments.

Therefore, using a large set of messages from text-based therapy session, we investigated if emotions extracted using a Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) based model trained on GoEmotions, a large dataset of Reddit posts annotated with 27 fine-grained emotions (Demszky et al., 2020), are stronger predictors of depression and anxiety status than counts of emotion-related word categories (LIWC). To this end, we first determined the association of each feature with the outcomes of interest in univariate regression analyses. Further, in order to provide clinical decision support to mental health practitioners, it is paramount to be able to classify previously unseen messages as indicating depression and/or anxiety. We therefore proceeded

	slope	R2
GoEmotions		
sadness	18.84 (16.50 - 21.18)**	0.782
admiration	-16.62 (-18.80 - -14.44)**	0.781
annoyance	12.61 (9.67 - 15.55)**	0.778
disappointment	19.01 (16.87 - 21.14)**	0.778
joy	-16.40 (-19.33 - -13.48)**	0.778
pride	-64.35 (-78.54 - -50.16)**	0.777
excitement	-28.34 (-33.18 - -23.49)**	0.777
disapproval	16.11 (12.99 - 19.23)**	0.776
approval	-7.81 (-9.27 - -6.36)**	0.776
confusion	9.65 (7.30 - 11.99)**	0.775
relief	-24.19 (-30.79 - -17.59)**	0.774
neutral	-0.83 (-1.60 - -0.06)*	0.774
anger	18.67 (14.38 - 22.97)**	0.774
disgust	29.79 (21.72 - 37.86)**	0.774
optimism	-6.15 (-8.28 - -4.03)**	0.773
realization	-1.08 (-2.74 - 0.59)	0.773
amusement	-10.96 (-14.85 - -7.07)**	0.772
fear	10.75 (7.44 - 14.06)**	0.771
nervousness	3.44 (0.84 - 6.05)*	0.771
caring	-2.77 (-6.03 - 0.49)	0.771
gratitude	-2.87 (-9.79 - 4.05)	0.771
embarrassment	11.85 (4.25 - 19.45)*	0.771
curiosity	0.03 (-2.56 - 2.62)	0.771
desire	2.08 (-1.10 - 5.26)	0.771
love	-1.96 (-5.22 - 1.31)	0.771
surprise	-4.00 (-10.18 - 2.18)	0.771
grief	134.76 (104.49 - 165.03)**	0.770
GoEmotions Ekman		
joy	-9.31 (-10.21 - -8.41)**	0.788
anger	18.53 (16.46 - 20.61)**	0.783
sadness	15.81 (14.06 - 17.56)**	0.779
disgust	48.43 (37.93 - 58.93)**	0.778
neutral	-0.11 (-1.17 - 0.96)	0.775
surprise	4.11 (2.47 - 5.75)**	0.774
fear	4.52 (2.25 - 6.80)**	0.772
LIWC		
sad	1.21 (1.02 - 1.40)**	0.781
i	0.25 (0.21 - 0.29)**	0.777
anger	0.84 (0.65 - 1.02)**	0.776
health	0.66 (0.52 - 0.80)**	0.775
anx	0.19 (0.05 - 0.34)*	0.774
we	-0.53 (-0.65 - -0.41)**	0.774
bio	0.41 (0.33 - 0.50)**	0.774

Table 1: PHQ-9 score univariate mixed-effects linear regression models coefficients and variance explained. * p<0.05. ** p<0.001

to train and evaluate a machine learning classifier using emotion features in conjunction with established depression-related LIWC features to predict depression and anxiety status in a held-out test set.

2 Methods

2.1 Data

We utilized a corpus of messaging therapy sessions from over 6,500 unique patients previously collected via the Talkspace platform (Hull et al., 2020). Talkspace offers a paid service utilizing licensed and credentialed therapists to conduct asynchronous, message-based therapy conversations. All patients and clinicians give written consent to the use of their data in a de-identified, aggregate format as part of the user agreement before they begin using the platform. Over the course of 12 weeks, patients engaged in two-way messaging therapy and completed depression questionnaires (9-item Patient Health Questionnaire, PHQ-9 (Kroenke et al., 2001)) as well as anxiety questionnaires (7-item General Anxiety Disorder questionnaire), every 3 weeks. For each available score, patient messages from the period in question ("(o)ver the last two (2) weeks") were concatenated into a single unit of analysis ("document"), resulting in up to 4 labeled data points per patient (weeks 3, 6, 9, and 12). All messages without a corresponding score were excluded from analysis. Data from baseline assessments were removed, as preliminary analysis suggested that messages before the week 0 mark introduced spurious associations due to differences between typical therapy dialog and the patient-therapist matching process, combined with generally worse symptom severity scores at the beginning of the study period. Participants were young (79% were 35 years old or younger), educated (75% had a Bachelor's degree or higher), and predominantly female (79%). Race and ethnicity were not systematically collected. There were over 13,000 text documents with both PHQ-9 and GAD-7 scores, totaling over 24 million words from over 337,000 messages. The original study was approved as exempt by the local institutional review board. The current study concerned secondary analysis of previously collected de-identified data, which is not considered human subjects research; nonetheless, data were stored on a secure server with study team member access only. All textual data were thoroughly de-identified by an automated algorithm before leaving their

	slope	R2
GoEmotions		
sadness	15.04 (12.96 - 17.12)**	0.728
admiration	-15.02 (-16.97 - -13.07)**	0.727
neutral	-1.00 (-1.72 - -0.29)*	0.725
joy	-16.99 (-19.53 - -14.44)**	0.724
approval	-6.80 (-8.14 - -5.47)**	0.724
fear	18.62 (15.32 - 21.93)**	0.724
annoyance	12.83 (10.20 - 15.46)**	0.724
excitement	-22.74 (-26.98 - -18.49)**	0.723
pride	-56.42 (-69.75 - -43.09)**	0.723
disappointment	14.05 (12.12 - 15.97)**	0.723
disapproval	12.97 (10.18 - 15.76)**	0.723
nervousness	11.91 (9.36 - 14.46)**	0.723
confusion	8.41 (6.33 - 10.48)**	0.721
anger	19.29 (15.38 - 23.19)**	0.721
relief	-22.16 (-28.05 - -16.28)**	0.720
optimism	-6.86 (-8.84 - -4.89)**	0.719
realization	-1.48 (-2.99 - 0.02)	0.718
amusement	-10.34 (-13.73 - -6.96)**	0.717
curiosity	-0.00 (-2.44 - 2.43)	0.717
caring	-1.94 (-5.11 - 1.24)	0.716
gratitude	-3.54 (-7.67 - 0.59)	0.716
desire	1.25 (-1.55 - 4.06)	0.716
love	-3.66 (-7.08 - -0.23)*	0.716
surprise	-6.42 (-11.87 - -0.97)*	0.716
embarrassment	10.33 (3.08 - 17.58)*	0.716
grief	118.01 (90.79 - 145.22)**	0.716
disgust	25.72 (18.68 - 32.76)**	0.715
GoEmotions Ekman		
joy	-8.62 (-9.42 - -7.82)**	0.736
anger	15.92 (14.08 - 17.76)**	0.727
disgust	44.78 (35.38 - 54.18)**	0.726
sadness	12.38 (10.83 - 13.93)**	0.725
neutral	-0.21 (-1.18 - 0.76)	0.722
fear	12.15 (9.97 - 14.33)**	0.722
surprise	3.27 (1.79 - 4.75)**	0.720
LIWC		
anx	0.73 (0.59 - 0.86)**	0.729
i	0.17 (0.13 - 0.21)**	0.726
sad	0.89 (0.72 - 1.05)**	0.726
anger	0.93 (0.76 - 1.10)**	0.724
we	-0.36 (-0.47 - -0.25)**	0.723
health	0.46 (0.33 - 0.59)**	0.717
bio	0.28 (0.21 - 0.36)**	0.716

Table 2: GAD-7 score univariate mixed-effects linear regression models coefficients and variance explained. * p<0.05. ** p<0.001

source, with all names, places, contact information, social media identifiers, and mentions of specific events removed.

LIWC 2015 was used to obtain the following word-count-based features: first-person singular pronouns ("I"), first-person plural pronouns ("we"), bio, health, sadness, anxiety, anger, positive emotion, and negative emotion. These features were selected on account of their track record of correlation with indicators of depression and anxiety in previous work (Tausczik and Pennebaker, 2010).

A BERT-based GoEmotions classifier pipeline using fine-tuned models available from the Hugging Face transformer library[1] was used to extract emotion features from each document. This model has been shown to approximate published results for performance in extracting emotions from the GoEmotions dataset (macro-average F1 score of ≈0.5 to ≈0.7, depending on the granularity of the emotions concerned). For further details of the training corpus and procedures used, we refer the reader to Demszky et al. (2020). After splitting documents into sentences and extracting emotions from the first 512 tokens of each sentence, scores were averaged over all sentences in a document to yield one set of emotion scores for the two-week period concerned. Only 38 of ≈13,000 documents contained sentences that were truncated due to being over 512 tokens long. The pipeline provides several output settings, resulting in different sets of emotions being extracted. Two sets of emotions were extracted. First, we extracted the set of 6 basic emotions proposed by Ekman (1992), consisting of sadness, joy, surprise, disgust, anger, fear, and a neutral category, which was assigned by annotators when they felt that no particular emotion was expressed. Second, we extracted the full set of 28 categories that were used to annotate the GoEmotions corpus, consisting of 27 fine-grained emotions described by Cowen and Keltner (2017), plus a neutral category. Finally, we calculated positive and negative emotion features by averaging the scores belonging to positive and negative emotions. The negative GoEmotions Ekman emotions are anger, disgust, fear, and sadness; joy is the only positive Ekman emotion. Negative fine-grained GoEmotions (Cowen) emotions encompass anger, annoyance, disappointment, disapproval, disgust, embarrassment, fear, grief, nervousness, remorse, and sadness. Admiration, amusement, approval, caring, desire, excitement, gratitude, joy, love, optimism, pride, and relief are the positive emotions in the fine-grained GoEmotions set. The interested reader is referred to Demszky et al. (2020) for further details on these groupings.

[1] https://github.com/monologg/GoEmotions-pytorch

2.2 Comparison of variables

A common approach to identifying associations of individual variables with an outcome of interest is to determine the statistical significance of the association between each candidate variable and the outcome by fitting univariate regressions. Linear regression models, however, require observations to be independent of each other. Because patients contribute between 1 and 4 observations in our dataset, this independence assumption is not met: two observations from the same patient may be expected to be more like each other than two observations from different patients. Mixed-effect linear regressions can be used to account for this. In such models, the within-patient and between-patient effects of the predictor variables on the outcome are separately accounted for. In other words, in addition to the "fixed effect" of the predictor variables on the outcome (the effect of interest), we model a "random effect" that is different for each patient, which is arbitrary but consistent across all observations for a given patient. In essence, the outcome is the linear combination of an emotion's global relationship to PHQ-9/GAD-7 scores and the patient-specific relationship of the emotion on scores (plus an intercept term for each effect as well as a residual error term). The univariate mixed-effect linear regression models for each emotion variable model the patient identity as a random effect and are of the following form:

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \gamma_{0i} + \gamma_{1i} X_{ij} + \epsilon_{ij}$$

Where Y_{ij} is the ith outcome (PHQ-9 score, GAD-7 score) for patient i, X_{ij} is the level of emotion in the jth document written by patient i, β_0 and β_1 are the fixed effect parameters (emotion), and γ_{0i} and γ_{1i} are the random effect parameters (patient ID), and ϵ_{ij} is the residual error for patient i's jth document. Models were fitted via Maximum Likelihood Estimation using the Statsmodels package for Python (Seabold and Perktold, 2010). Statsmodels calculates p-values using t-tests. We report the explanatory power of each feature as the amount of variance explained (R2).

Following a similar process, we fitted bivariate mixed-effects models using the positive and negative emotion variables from each feature source.

2.3 Prediction

Next, using the Scikit-Learn package for Python (Pedregosa et al., 2011), we trained random forest classifiers to predict binary depression (MDD) and anxiety (GAD) status from 49 features: 7 Ekman emotion categories from GoEmotion, as well as the positive and negative emotion variables calculated from Ekman emotions; 27 fine-grained emotions plus neutral, as well as the positive, and negative emotion variables calculated from the 27 fine-grained emotions; 5 LIWC emotion variables (positive emotion, negative emotion, anxiety, anger, sadness); and 4 LIWC variables with an established relationship to depression (I, we, biology, health) (Rude et al., 2004; De Choudhury et al., 2013b; Eichstaedt et al., 2018; Sonnenschein et al., 2018; Burkhardt et al., 2021). We first trained random forest classifiers using each individual feature set. Then, we trained models using combinations of these feature sets to evaluate their relative contribution (LIWC non-emotion variables combined with each set of emotion variables from the three sources). Then, we trained another random forest classifier on all available features. For this model, relative feature importance was calculated using SHAP (Lundberg and Lee, 2017).

To avoid information leakage due to within-patient effects (Saeb et al., 2017), data were split into training and test sets such that all observations from an individual patient were kept within the same fold. Patients were assigned to the training (80%) and test (20%) populations, resulting in a training set of 4,913 patients (with 10,006 observations) and a test set of 1,638 patients (with 3,321 observations). Average PHQ-9 across all observations did not significantly differ between training and test observations.

Hyperparameters (number of estimators, maximum number of features, maximum tree depth, minimum number of samples for splitting, minimum number of samples per leaf, using or not using bootstrap) were automatically selected (based only on the training data) via 3-fold cross-validation, a process where, for each hyperparameter combination, each of the three folds is held out in turn, while a model is trained on the remaining 2 folds; this way, 3 scores are produced per hyperparameter combination, and their average represents the score for that hyperparameter set. Finally, the hyperparameters that produced the best score are selected, and a final model with those hyperparameters is trained on all training data, then tested on the held-out test set.

A binary prediction target was used to align predictions with the clinical task of classifying a diagnosis as present or absent. A cut-off between 8 and 11 was previously found to have a clinically acceptable tradeoff between sensitivity and specificity when dichotomizing PHQ-9 scores for diagnosis of major depressive disorder (MDD) (Manea et al., 2012). Therefore, we considered a PHQ-9 score of 10 or more (depression severity of moderate, moderately severe, or severe) as indicating MDD for the purposes of this work. A PHQ-9 score of 9 or less (depression severity of mild or none) was considered non-depressed. As the GAD-7 has been found to have acceptable properties for identification of generalized anxiety disorder (GAD) at a cutoff of 7-10 (Plummer et al., 2016; Spitzer et al., 2006), a GAD-7 score of 10 or more was considered an indicator of GAD, and a score of 9 or less was considered an indicator of a negative diagnosis for this condition.

3 Results

3.1 Comparison of variables

The variance in PHQ-9 and GAD-7 scores, respectively, explained by each individual emotion variable and by variable pairs is shown in Figure 1, Table 1, and Table 2. Emotion variables that were obtainable from all three feature sources were anger and sadness as well as the summary dimensions of positive and negative emotion. With BERT-based models, these are composites of individual predictions returned by the model, while LIWC returns a summary value as an individual feature. The variance in PHQ-9 scores explained by these directly comparable variables is shown in Figure 1, along with the variance explained by the combination of positive and negative emotion features. The three feature extraction approaches resulted in features that explained similar portions of the variance; LIWC explained slightly more, except for anger and sadness, where the GoEmotions Ekman and GoEmotions Cowan variables explained more, respectively. The GoEmotions Cowan variable for sadness was more explanatory than the GoEmotions Ekman variable, but the Ekman anger variable outperformed the fine-grained anger vari-

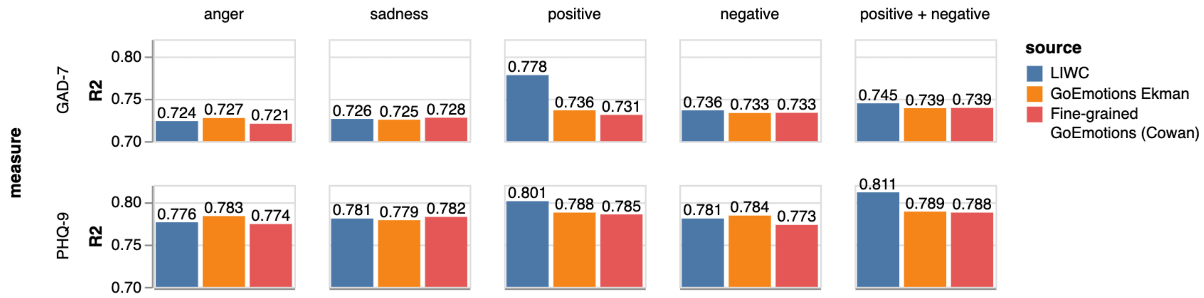

Figure 1: PHQ-9 and GAD-7 score variance explained by comparable features from LIWC, GoEmotions (Ekman set), and GoEmotions (fine-grained set)

able. The combination of positive and negative emotion explained more variance than either positive or negative emotion alone, except when LIWC positive emotion was assessed for GAD-7. Notably, LIWC's positive emotion variable appears to be more explanatory than anger, sadness, and negative emotion for both PHQ-9 and GAD-7, and even the combination of positive and negative emotion for GAD-7.

All individual emotions, as quantified by each of the three feature extraction approaches, are shown in Table 1. Expressions of realization, caring, gratitude, curiosity and desire were not significantly associated with either anxiety or depression. Love and surprise were not predictive of depression, but were associated with anxiety. Both were significantly associated with sadness, fear, and nervousness; however, sadness was more strongly related to depression, and fear and nervousness were more strongly related to anxiety. Joy was roughly equally associated with anxiety and depression, across both GoEmotions feature sets. The emotions with the largest differences between more and less depressed individuals were grief, pride, excitement, relief, and disgust. The emotions with the largest differences with respect to anxiety were grief, disapproval, approval, relief, and disgust.

3.2 Prediction

In contrast to the multivariate models, results from predictive modeling experiments show a clear advantage for deep learning models, with the best overall performance by ROC and F1 score achieved using GoEmotions Cowen features for both MDD and GAD. As shown in Table 3, the models including only the non-emotion LIWC features achieved an area under the receiver-operator characteristic curve (AUROC) of 0.577 for MDD and 0.549 for GAD. When using emotion features only, the fine-grained GoEmotions set performed best. For both MDD and GAD, adding LIWC emotion features to LIWC non-emotion features improved predictive performance less than adding GoEmotions Ekman features, which improved the model less than adding the fine-grained GoEmotions set. Using all emotion features concurrently ("all three") slightly improved performance for both GAD and MDD (by F1 score but not ROC in the latter case).

The relative importance of all features for the MDD and GAD models is shown in Table 4. Fear was ranked higher for predicting GAD than for predicting MDD. Sadness was ranked higher for predicting MDD than for predicting GAD.

4 Discussion

In this work, we showed that neural network models such as the BERT-based GoEmotions classifier can outperform LIWC, a straightforward, broadly adopted word-counting method for extracting emotion features from natural language. We further confirmed that some emotions not traditionally associated with depression and anxiety can be predictive of these diagnoses; specifically, pride. Finally, we showed that using LIWC features together with emotion features derived using GoEmotions predict depression/anxiety status with reasonable accuracy. This finding is important, in that further development of such tools could lead to better detection of emotional change during treatment in a way that could be derived naturally in the client/clinician encounter. NLP applied to such naturalistic data has been used for measuring clinician skills in delivering psychotherapy with some success (Flemotomos et al., 2021); here, rather than using such tools for quality measurement, linguistic analysis of affect could be used to detect depression/anxiety severity and client response to treatment.

Both LIWC variables and GoEmotions variables

	ROC	F1	Pr	Rc
MDD				
LIWC non-emo	0.577	0.413	0.525	0.341
LIWC emo	0.621	0.471	0.561	0.405
GoEmo Ekman	0.643	0.493	0.583	0.427
GoEmo	0.662	0.522	0.613	0.455
LIWC non-emo +				
LIWC emo	0.640	0.484	0.569	0.420
GoEmo Ekman	0.655	0.498	0.585	0.434
GoEmo	0.671	0.514	0.615	0.441
All three	0.671	0.520	0.612	0.453
GAD				
LIWC non-emo	0.549	0.290	0.478	0.209
LIWC emo	0.613	0.405	0.541	0.324
GoEmo Ekman	0.643	0.443	0.550	0.371
GoEmo	0.652	0.444	0.565	0.366
LIWC non-emo +				
LIWC emo	0.617	0.401	0.529	0.324
GoEmo Ekman	0.637	0.441	0.548	0.369
GoEmo	0.654	0.451	0.568	0.374
All three	0.657	0.456	0.567	0.382

Table 3: AUROCs, F1 score (positive class), precision, and recall of random forest model trained with just the non-emotion LIWC features, and trained with the non-emotion LIWC features plus LIWC emotion, GoEmotion Ekman and the full GoEmotion feature sets, for predicting MDD (PHQ-9 score $\geq$ 10) and GAD (GAD-7 score $\geq$ 10).

explained a large portion of the variance in univariate mixed-effect regressions: R2 values ranged from 0.770 to 0.788 when modeling PHQ-9 scores as outcome, and from 0.715 to 0.736 when modeling GAD-7 scores as outcome. Therefore, LIWC and GoEmotions features both capture valuable information. GoEmotions features marginally outperformed 2 out of 4 of the equivalent LIWC features for predicting GAD-7 and 3 ouf of 4 features for predicting PHQ-9. For predicting binary depression (MDD) and anxiety (GAD) status, the emotion set resulting in the best predictive performance when combined with LIWC's non-emotion features was the full GoEmotions set.

However, despite the availability of pre-trained models, neural networks can have high computational demands. Consequently, using BERT-based models may not be justified if the cost of model inference outweighs the potential benefits. Therefore, the decision to include these features should be evaluated for each individual predictive analytics project and dataset, weighing the added predictive performance observed at development time with the costs to include the features in production (e.g. a deployed clinical decision support tool continuously evaluating patient-generated messages in real-time), given the available compute resources. Similarly, on-device processing to preserve data privacy can be accomplished with LIWC (Liu et al., 2022), but doing this with a BERT-based model would challenge some contemporary and most legacy smartphone devices.

Depression affects individuals in many ways and expresses itself in various behavioral and thought patterns that may not be fully captured with the high-level categories of positive and negative affect. GoEmotions' main strength therefore lies in its ability to extract fine-grained features spanning the breadth of human emotion, capturing depressed individuals' emotional experiences comprehensively. The different emotion feature sets appeared to be somewhat complementary, as evidenced by the additive performance metrics shown in Table 3; however, when predicting depression, the combination of non-emotion LIWC features and fine-grained GoEmotions features was as predictive as all features combined, suggesting that all signal is contained within this feature subset. In this work, this breadth enabled us to delineate differences in how different types of emotions are associated with depression and anxiety.

Depression severity was associated with large differences in grief, pride, excitement, relief, and disgust. In agreement with generally lower reactivity (Rottenberg, 2017), less excitement was predictive of depression. Grief manifestations are similar to depression symptoms; though grief in itself is not pathological, it often co-occurs with depression (Aoyama et al., 2018). Additionally, depressed individuals expressing less pride than their non-depressed counterparts might be expected on account of lower self-image, and matches findings presented by Gruber et al. (2011). Caused by a perception of violations of moral and social norms, internally directed disgust, also termed self-disgust or self-loathing, has been reported to be associated with both depression and anxiety symptoms (Ille et al., 2014). We further found that increased disapproval - and conversely, decreased approval - were associated with anxiety symptoms. This may be explained by disturbances in interpersonal sensitivity and an inclination to be self-critical, which have been described as characteristic of anxiety

	MDD	GAD
1	LIWC we	GE negemo
2	GEE posemo	GEE negemo
3	GEE joy	GEE joy
4	GEE sadness	GEE posemo
5	GE negemo	GE posemo
6	LIWC bio	LIWC bio
7	GE disappointment	GE fear
8	GEE negemo	GE sadness
9	LIWC sad	LIWC health
10	LIWC i	LIWC we
11	LIWC health	LIWC posemo
12	GE posemo	GE realization
13	GE excitement	GEE sadness
14	GE admiration	GE nervousness
15	GE sadness	GEE fear
16	GEE anger	LIWC negemo
17	GE confusion	GE pride
18	GE pride	LIWC anx
19	GE disapproval	GE joy
20	GE joy	LIWC i
21	GEE disgust	GE disappointment
22	GE realization	GE admiration
23	LIWC posemo	GEE anger
24	GE relief	GE excitement
25	GE approval	GE disgust
26	GE disgust	GE confusion
27	LIWC negemo	GEE disgust
28	GE grief	GE grief
29	GEE fear	GE neutral
30	GEE neutral	GE relief
31	GE fear	GEE neutral
32	GE desire	GEE surprise
33	GE remorse	LIWC sad
34	GE curiosity	GE desire
35	GE nervousness	GE neutremo
36	GE embarrassment	GE curiosity
37	LIWC anx	GE gratitude
38	GE optimism	GE disapproval
39	GE amusement	GE love
40	GE neutremo	GE embarrassment
41	GE neutral	GE anger
42	GE gratitude	GE approval
43	GE love	GE annoyance
44	GE annoyance	GE amusement
45	GE surprise	GE remorse
46	GEE surprise	GE caring
47	LIWC anger	GE surprise
48	GE caring	GE optimism
49	GE anger	LIWC anger

Table 4: Random forest classifier features in order of importance (most important first) for predicting MDD and GAD, as calculated by SHAP (Lundberg and Lee, 2017). GE = GoEmotions. GEE = GoEmotions Ekman

(Ille et al., 2014).

Non-emotion LIWC features have established utility for predicting depression and anxiety. These features capture aspects of symptomatology outside emotion, such as increased self-focus, social isolation, and usage of health-related words. Non-emotion LIWC features would therefore be expected to be complementary to emotion features, and our work confirms that and leveraging both may achieve the best results. We trained a machine learning model using these features in conjunction with emotion features to predict depression (AUROC 0.671) and anxiety (AUROC 0.657). That these models show similar performance using the same features to predict different outcomes may be explained by the large overlap in symptoms between anxiety and depression, e.g. both are characterized by negative self-talk and hopelessness. Additionally, depression and anxiety are often comorbid; indeed, in this dataset, 74.5% of assessments with a GAD-7 score above the diagnosis threshold also had a positive depression finding, and 70.6% of positive anxiety questionnaires also had a positive anxiety finding.

There are important ethical considerations when analyzing patient-generated natural language to infer mental state. Any passive monitoring of patient-generated data may be considered invasive. Due to the sensitive nature of personal health data, such data are subject to protections that do not apply to non-health data. When health-related insights are derived from data that may be neither private nor health-related (e.g. social media posts), obtaining informed consent and handling inferences with appropriate care is paramount. While academic studies such as the current work are governed by rigorous institutional ethics guidelines regarding consent and data sharing, different rules apply to healthcare organizations and commercial entities. The use of technologies such as the ones presented here may be acceptable if conducted by trusted entities, such as healthcare providers, in order to support care (Areán et al., 2021); on the other hand, consumers may be wary of commercial entities conducting such analyses. Further research, as well as applications of the findings presented here, must take such considerations into account.

This work has several limitations. The data used here stem from predominantly female, young, and well-educated participants, and results may therefore not generalize to populations with a different makeup. If predictive algorithms were to be deployed in practice, fairness may be a concern if predictive performance differs for underrepresented groups. In addition, the GoEmotions dataset used to train the BERT-based models is drawn from Reddit, which has been shown to have a disproportionately high representation of young male users (Duggan and Smith, 2013). Though it is encouraging that models trained on these data produce features that correlate well with symptom severity in the current study, the development of annotated datasets drawn from a more diverse population may lead to models that better address linguistic and cultural differences in the ways in which emotions are expressed.

Several features used in the random forest classification model are expected to be highly redundant (e.g. GoEmotions Cower sadness, GoEmotions Ekman sadness, and LIWC sadness; calculated negative emotion variables which are calculated using sadness). However, interdependent features should not affect the random forest's ability to leverage all features optimally to optimize predictive performance.

This work enables and informs future work. We showed that BERT-based emotion features are associated with depression and anxiety status; however, this work did not assess longitudinally if changes in emotion track with changes in depression and anxiety. While existing work demonstrated this relationship for depression-related LIWC features (Burkhardt et al., 2021), future work may aim to ascertain whether changes in emotion features over time also predict longitudinal patient trajectories. This work also informs feature selection for future work in depression and anxiety prediction. Emotion variables can be obtained with a range of extraction approaches. Our results indicate the GoEmotions variables may be a better choice than LIWC for emotions. Nevertheless, LIWC features have a place in future work. LIWC's syntactic and topic features were shown in prior work to be associated with depression scores as well as longitudinal patient trajectories and continued to demonstrate utility in this work.

We determined that fine-grained emotions measured in the language of individuals are associated with and predict anxiety and depression status. The associations we found reflect previous findings. This work thus contributes evidence of the reliability of such measurement approaches, supporting the use of these methods in future work investigating the nature of depression and anxiety. For example, these features could aid investigations into depression phenotypes through cluster analysis, as well as psychology research investigating the differential expression of similarly-valenced emotions in depression and anxiety, e.g. by aiding data collection.

Additionally, this work has important clinical implications. Measurement-based care is facilitated by periodic progress assessments, but additional data collection incurs additional workload. In text-based therapy, depression and anxiety status may instead be automatically determined from already-available patient messages. In clinical settings, interpretability is essential; thus, models based on interpretable features such as emotions may be preferred over black-box models classifying raw text directly. Future work may therefore investigate opportunities to leverage emotion-based predictive models for clinical decision support.

5 Conclusion

Extraction methods differ in the quality of emotion features extracted. With the data and approaches presented here, emotion features extracted by the GoEmotions BERT-based model not only explained more variance in univariate mixed-effect regressions, but also contributed significantly to predictions of depression and anxiety status by a random forest classifier. Further, while non-emotion variables obtained from LIWC remain valuable in linguistic modeling tasks, GoEmotions' level of granularity offers clinically relevant nuance that prevailing tools cannot capture.

5.1 Acknowledgments

This work was supported by the National Library of Medicine (grant number 67-3780) and by Innovation Grant "Informatics-Supported Authorship for Caring Contacts (ISACC)" from the Garvey Institute for Brain Health Solutions.

5.2 Conflicts of Interest

TDH is an employee of the platform that provided the data.

References

Ananda Amstadter. 2008. Emotion regulation and anxiety disorders. *Journal of anxiety disorders*, 22(2):211–221.

Maho Aoyama, Yukihiro Sakaguchi, Tatsuya Morita, Asao Ogawa, Daisuke Fujisawa, Yoshiyuki Kizawa, Satoru Tsuneto, Yasuo Shima, and Mitsunori Miyashita. 2018. Factors associated with possible complicated grief and major depressive disorders. *Psycho-Oncology*, 27(3):915–921.

Patricia A Areán, Abhishek Pratap, Honor Hsin, Tierney K Huppert, Karin E Hendricks, Patrick J Heagerty, Trevor Cohen, Courtney Bagge, and Katherine Anne Comtois. 2021. Perceived Utility and Characterization of Personal Google Search Histories to Detect Data Patterns Proximal to a Suicide Attempt in Individuals Who Previously Attempted Suicide: Pilot Cohort Study. *Journal of Medical Internet Research*, 23(5):e27918.

Aaron T. Beck. 1967. *Depression: clinical, experimental, and theoretical aspects*. Hoeber Medical Division, Harper & Row, New York.

Hannah A. Burkhardt, George S. Alexopoulos, Michael D. Pullmann, Thomas D. Hull, Patricia A. Areán, and Trevor Cohen. 2021. Behavioral Activation and Depression Symptomatology: Longitudinal Assessment of Linguistic Indicators in Text-Based Therapy Sessions. *Journal of Medical Internet Research*, 23(7):e28244.

Stevie Chancellor and Munmun De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: a critical review. *npj Digital Medicine*, 3(1).

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying Mental Health Signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60, Stroudsburg, PA, USA. Association for Computational Linguistics.

Alan S. Cowen and Dacher Keltner. 2017. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences of the United States of America*, 114(38):E7900–E7909.

Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013a. Social media as a measurement tool of depression in populations. In *Proceedings of the 5th Annual ACM Web Science Conference on - WebSci '13*, pages 47–56, New York, New York, USA. ACM Press.

Munmun De Choudhury, Scott Counts, Eric J. Horvitz, and Aaron Hoff. 2014. Characterizing and predicting postpartum depression from shared facebook data. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*, pages 625–637.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013b. Predicting Depression via Social Media. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, pages 128–137.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A Dataset of Fine-Grained Emotions.

Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1(Mlm):4171–4186.

Maeve Duggan and Aaron Smith. 2013. 6% of Online Adults are reddit Users. *Pew Research Center*, pages 1–10.

Johannes C. Eichstaedt, Robert J. Smith, Raina M. Merchant, Lyle H. Ungar, Patrick Crutchley, Daniel Preoţiuc-Pietro, David A. Asch, and H. Andrew Schwartz. 2018. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences of the United States of America*, 115(44):11203–11208.

Paul Ekman. 1992. Are There Basic Emotions? *Psychological Review*, 99(3):550–553.

Nikolaos Flemotomos, Victor R. Martinez, Zhuohao Chen, Karan Singla, Victor Ardulov, Raghuveer Peri, Derek D. Caperton, James Gibson, Michael J. Tanana, Panayiotis Georgiou, Jake Van Epps, Sarah P. Lord, Tad Hirsch, Zac E. Imel, David C. Atkins, and Shrikanth Narayanan. 2021. Automated evaluation of psychotherapy skills using speech and language technologies. *Behavior Research Methods*, pages 690–711.

June Gruber, Christopher Oveis, Dacher Keltner, and Sheri L. Johnson. 2011. A discrete emotions approach to positive emotion disturbance in depression. *Cognition and Emotion*, 25(1):40–52.

Sharath Chandra Guntuku, David B. Yaden, Margaret L. Kern, Lyle H. Ungar, and Johannes C. Eichstaedt. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, 18:43–49.

Thomas D. Hull, Matteo Malgaroli, Philippa S. Connolly, Seth Feuerstein, and Naomi M. Simon. 2020. Two-way messaging therapy for depression and anxiety: longitudinal response trajectories. *BMC Psychiatry*, 20(1):297.

Rottraut Ille, Helmut Schöggl, Hans Peter Kapfhammer, Martin Arendasy, Markus Sommer, and Anne Schienle. 2014. Self-disgust in mental disorders - Symptom-related or disorder-specific? *Comprehensive Psychiatry*, 55(4):938–943.

Bennett Kleinberg, Isabelle van der Vegt, and Maximilian Mozes. 2020. Measuring emotions in the covid-19 real world worry dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*.

Kurt Kroenke, Robert L Spitzer, and Janet B W Williams. 2001. The PHQ-9. *Journal of General Internal Medicine*, 16(9):606–613.

Tony Liu, Jonah Meyerhoff, Johannes C Eichstaedt, Chris J Karr, Susan M Kaiser, Konrad P Kording, David C Mohr, and Lyle H Ungar. 2022. The relationship between text message sentiment and self-reported depression. *Journal of affective disorders*, 302:7–14.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

Laura Manea, Simon Gilbody, and Dean McMillan. 2012. Optimal cut-off score for diagnosing depression with the Patient Health Questionnaire (PHQ-9): a meta-analysis. *Canadian Medical Association Journal*, 184(3):E191–E196.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, and E. Cournapeau, D. and Brucher, M. and Perrot, M. and Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

James W. Pennebaker, R.J. Booth, and M.E. Francis. 2007. Linguistic Inquiry and Word Count: LIWC.

Faye Plummer, Laura Manea, Dominic Trepel, and Dean McMillan. 2016. Screening for anxiety disorders with the gad-7 and gad-2: a systematic review and diagnostic metaanalysis. *General hospital psychiatry*, 39:24–31.

Jonathan Rottenberg. 2017. Emotions in Depression: What Do We Really Know? *Annual Review of Clinical Psychology*, 13:241–263.

Stephanie S. Rude, Eva Maria Gortner, and James W. Pennebaker. 2004. Language use of depressed and depression-vulnerable college students. *Cognition and Emotion*, 18(8):1121–1133.

Sohrab Saeb, Luca Lonini, Arun Jayaraman, David C. Mohr, and Konrad P. Kording. 2017. The need to approximate the use-case in clinical machine learning. *GigaScience*, 6(5):1–9.

Kelli Scott and Cara C. Lewis. 2015. Using measurement-based care to enhance any treatment. *Cognitive and Behavioral Practice*, 22(1):49–59.

Skipper Seabold and Josef Perktold. 2010. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.

Judy Hanwen Shen and Frank Rudzicz. 2017. Detecting anxiety through reddit. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology—From Linguistic Signal to Clinical Reality*, pages 58–65.

Anke R. Sonnenschein, Stefan G. Hofmann, Tobias Ziegelmayer, and Wolfgang Lutz. 2018. Linguistic analysis of patients with mood and anxiety disorders during cognitive behavioral therapy. *Cognitive Behaviour Therapy*, 47(4):315–327.

Robert L. Spitzer, Kurt Kroenke, Janet B.W. Williams, and Bernd Löwe. 2006. A brief measure for assessing generalized anxiety disorder: The GAD-7. *Archives of Internal Medicine*, 166(10):1092–1097.

Yla R. Tausczik and James W. Pennebaker. 2010. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, 29(1):24–54.

Shannon Wiltsey Stirman and James W. Pennebaker. 2001. Word Use in the Poetry of Suicidal and Nonsuicidal Poets. *Psychosomatic Medicine*, 63(4):517–522.

Benjamin Zablotsky and Emily P. Terlizzi. 2020. Mental Health Treatment Among Adults: United States, 2019. *NCHS data brief*, (380):1–8.