Predicting Borderline Personality Disorder Features from Free Response Text with Machine Learning and Natural Language Processing Techniques

Poster Number: P6-070

Eric Lin MD, Sarah K Fineberg MD PhD

Yale Department of Psychiatry, Yale University School of Medicine, New Haven, CT

Introduction

- Borderline personality disorder (BPD) patients are high utilizers of care and are at high risk of self-harm and suicide, but current risk assessment methods are poorly predictive of actual risk.
- Digital phenotyping can be used to evaluate large datasets derived from patients' active and passive interaction with technology, with the goal of making personalized precision predictions regarding such health risks.
- One example of digital phenotyping is digital phenotyping. Unstructured social media text has been used to make reproducible and accurate predictions about personality traits of individuals.

Hypothesis:
Language analytic tools can predict self-reported BPD symptom level in a large online community sample.

Methods

Amazon Mechanical Turk (MTurk) dataset:

- Mturk is an online labor market website that has been used in psychological research to gather large datasets.
- This dataset is from another study that examined prevalence rates of psychopathology in an Mturk sample.
- Exclusion criteria: schizophrenia, TBI, learning disability, those who failed attention checks, who were extreme responders (ie. all "5"s on all Likert scales) or completed the surveys too quickly
- $0.40 task to complete a medical history questionnaire, multiple clinical rating scales, and a 500 character minimum writing sample with the prompt "Tell us about yourself."
- Scales scores and diagnoses were not confirmed with gold standard in-person interviews

Creating Classes for Analyses:

- Individuals are separated based on their scores for the online test with the Structured Clinical Interview for DSM-IV Personality Questionnaire for BPD (SCID-II PQ BPD). The three classes described below were used to split the entire dataset, but only the first two are used for the binary classification task.
- "Current BPD Features" group is defined as those meeting 5+ out of 9 BPD criteria within the past two years.
- "Never BPD features" group is defined as those meeting < 5 BPD criteria in their lifetimes.
- Another group identified in the referenced study as "Remitted BPD features" was defined as meeting 5+ BPD criteria in their lifetimes but less than 5 in the past 2 years.

Text Tokenization and Embedding:

- Punctuation is removed from text, and all words are made lower case.
- Embedding techniques (approaches to turn words into computable values):
  - Bag of words term frequency (TF): each unique word across all the text samples are assigned a unique identifier. Writing samples are represented as an unordered (all syntax is lost) set of word counts.
  - Bag of words term frequency - inverse document frequency (TF-IDF): same as TF except that words are inversely weighted to their frequency in all of the text samples. One intended benefit is to weigh frequent words (such as "the") less heavily than less common words.
  - Paragraph average Word2Vec (w2v): neural network based method to embed words. As opposed to prior embedding techniques, w2v captures a word's semantic meaning. Text samples are represented as the average vector of the paragraphs' words' w2v vector representations.

Machine Learning Classifiers:

- The data is separated into an 80/20 test train split. Stratified 5 fold cross validation was checked within the training data.
- Supervised classifiers: logistic regression (LR), naïve bayes (NB), support vector machines (SVM), random forests (RF), and convolutional neural networks (CNN)
- Scores: precision (aka positive predictive value), recall (aka sensitivity), F1 (harmonic mean of precision and recall), and test set accuracy ((the algorithm's rate in correctly guessing the true class)
  - Test set prediction accuracy is the primary metric

Table 1. 5-fold stratified cross validation average F1, precision, recall, F1 score, and test set accuracy are reported. The best performing model overall in terms of test set accuracy is bolded. Convolutional neural network model results are not shown given the inconsistency in reported metrics.

Model and Embedding	CV Average F1	Precision	Recall	F1	**Accuracy**
LR TF	0.626	0.624	0.623	0.623	0.623
LR TF-IDF	0.646	0.64	0.64	0.64	0.64
LR word2vec	**0.65**	**0.696**	**0.695**	**0.695**	**0.695**
Multinomial NB TF	0.66	0.642	0.642	0.642	0.642
Multinomial NB TF-IDF	0.65	0.655	0.644	0.637	0.644
Complement NB TF	0.663	0.639	0.637	0.636	0.637
Complement NB TF-IDF	0.638	0.648	0.633	0.622	0.632
Bernoulli NB TF	0.648	0.671	0.663	0.658	0.662
Bernoulli NB TF-IDF	0.648	0.671	0.663	0.658	0.662
Linear SVC TF	0.623	0.612	0.612	0.612	0.612
Linear SVC TF-IDF	0.646	0.633	0.633	0.633	0.633
Linear SVC word2vec	0.65	0.693	0.693	0.693	0.693
Random Forest TF	0.646	0.66	0.649	0.642	0.648
Random Forest TF-IDF	0.643	0.624	0.619	0.614	0.618
Random Forest word2vec	0.638	0.62	0.613	0.614	0.609

Figure 1. Using the LIME package (described in the figure below), the fifteen most relevant features contributing to predicting each class for the word2vec logistic regression algorithm (best performing model) are shown.

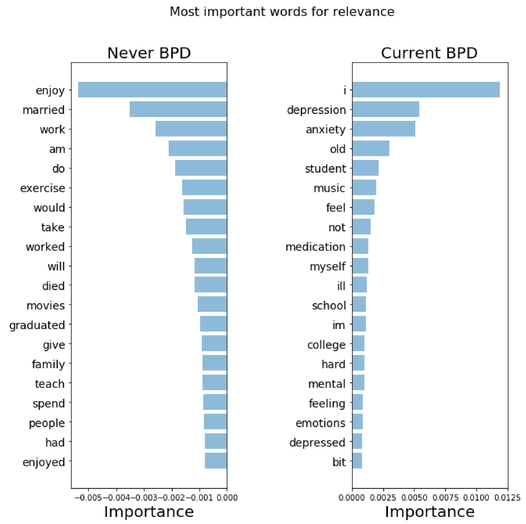
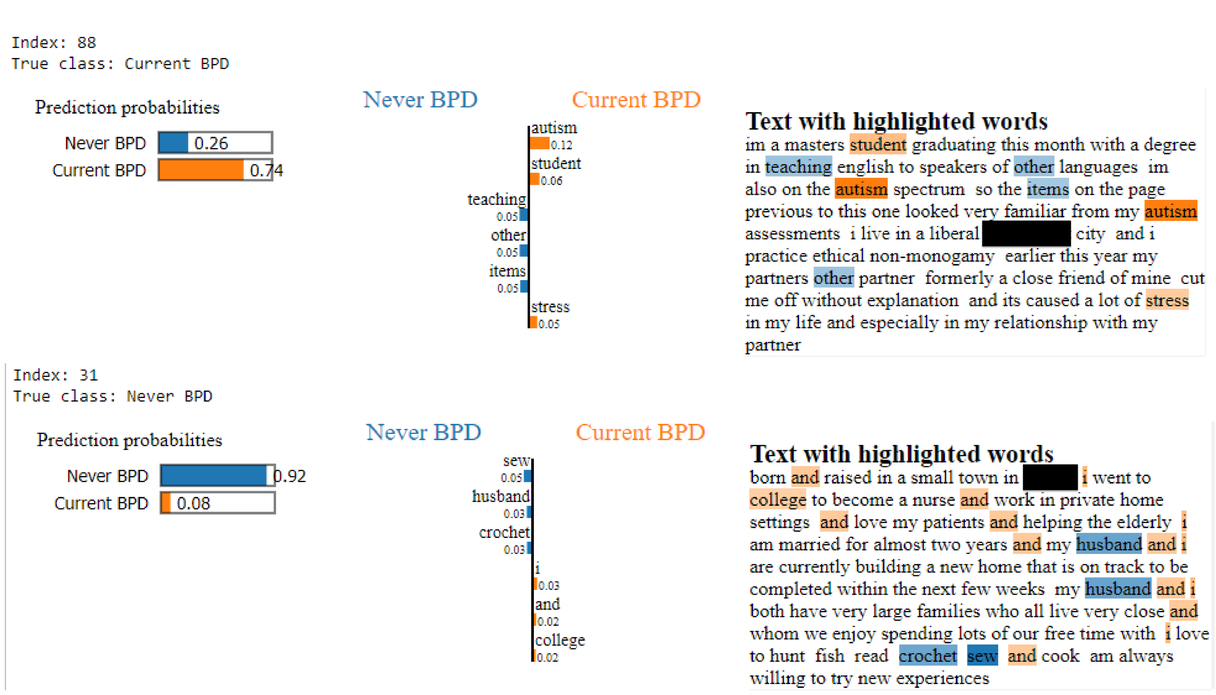


Figure 2. Example instances of predictions using the word2vec logistic regression algorithm are shown. The true class label is shown in the upper left corner, and the algorithm's prediction probabilities are shown just below. Using the LIME package, the three most important features for predicting either outcome in the individual text sample are highlighted. LIME is a machine learning explanation library in Python; it determines the importance of features by perturbing the input to the algorithm and observing the subsequent change to the individual prediction.

Two examples of correct predictions with high confidence are shown here.



Results

Population Sample Characteristics:

- Of 3,633 initial participants, 512 were excluded.
- "Current BPD features" class n = 1,020
- "Never BPD features" class n = 1,136
- "Remitted BPD features" class n = 865
- Only the first two classes were used for the study because an algorithm could not be trained to have any reasonable level of predictive accuracy (prediction accuracy was often close to 50%).
- Significant differences ($p < 0.001$ unless otherwise specified) by Pearson Chi-square values across the three classes were noted:
  - Diagnoses: BPD, depression, anxiety, bipolar disorder, substance use disorder, and chronic pain ($p = 0.002$)
  - Medications: antidepressants, mood stabilizers, antipsychotics, benzodiazepines, and other psychotropics ($p = 0.002$)

Text Sample Characteristics:

- Total number of unique words: 10,937
- Maximum length of any text sample: 344 words

Algorithm Results:

- **Logistic regression classifier with averaged word2vec vectors produced the highest accuracy as shown in Table 1.**
- This model also correctly predicted the "Current BPD features" class most often (data not shown) despite having lower cross-validation scores.
- Repeated efforts to train the CNN could not reliably achieve an accuracy much higher than 53%.

Discussion

- Preliminary results for binary classification of low ("Never BPD features" class) versus high BPD features ("Current BPD features" class) are promising.
- According to LIME, the algorithm seems to emphasize comorbid conditions, disease severity, and demographic factors to make its predictions. Of note, many of these can correspond to significant differences between classes in the sample characteristics.
- Despite various hyperparameter adjustments to the CNN, the CNN repeatedly overfit (test set accuracy scores significantly lower than training set accuracy) which suggests that the dataset may be too small for deep learning methods.
- Newer deep learning architectures, embeddings, and language models for natural language processing may improve classification even at this sample size.
- Automatic algorithms hold promise to predict borderline personality disorder features at an individual level, potentially offering opportunities for early intervention or for monitoring treatment response.

Future Directions

- Trial newer embedding models, language models, and deep learning classifiers
- Use demographic and medical features for the algorithm to evaluate in addition to text samples
- Have psychiatrists label text samples and compare against the algorithm
- Redo analyses with other clinical ratings scales in the dataset
- Redo analyses for less complex or more specific and easily identifiable phenomena such as chronic suicidality (as opposed to a DSM disorder)