# Predicting outcomes in online chatbot-mediated therapy

David S. Lim
dslim@ stanford.edu
CS229 Fall 2017

Stanford ENGINEERING
Computer Science

Stanford ENGINEERING
Computer Science

## Background / Problem

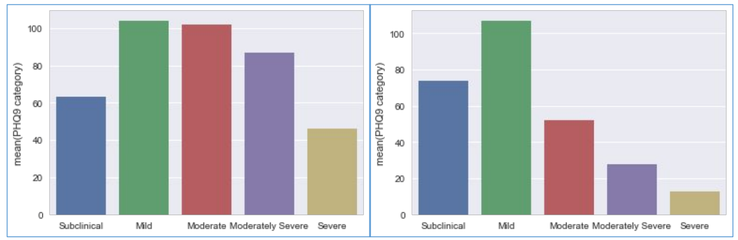
- This project aims to predict the outcome of 2 weeks of brief online therapy with the Woebot mental health chatbot
- Outcome class was measured using PHQ-9 scale (1-27) for depression through a survey at 2 weeks
- Input data:
  - User Data: baseline user survey data and usage stats
  - Mood Data: text data from daily Woebot check-in
  - All Text: all user inputs into the Woebot chatbot
- Predicting outcomes could allow for online chatbot therapy to identify to non-responders early and adapt program to help users without surveys like a human therapist

## Dataset

- User transcripts & survey data from clinical study on Woebot. 274 users made it to 2 week timepoint
- User data: patient history, demographics, engagement metrics (daily use, text per message)
- Mood data: 3948 user checkins
- All Text: 13,436 user messages to Woebot in 2 week span


## Features

- User data: 10 features chosen from 47 features per user
- Mood data
  - 25 topics extracted through LDA clustering
  - Features are vector of 25 topic probabilities per user
- All Text:
  - Top 100 most frequent words in a TF-IDF matrix




Baseline PHQ-9 After 2 weeks PHQ-9

**Figure 1: Change in distribution of depression severity over two weeks**

Topic #0: meh sure okay head nauseous ambivalent awful
Topic #1: good pretty feel optimistic energetic stable trying
Topic #2: exhausted accomplished sore lost high stress unsure thoughtful
Topic #3: hungry normal indifferent conflicted refreshed introspective chilled
Topic #4: frustrated scared kind upset pain terrible distracted try
Topic #5: feel stressed worried work day like im time
Topic #6: fine annoyed chill guilty focused week mixed past
Topic #7: sleep hurts need determined ill inspired heavy rushed
Topic #8: bit confused blah long flat weird grumpy sorry….
Topic #9: overwhelmed slightly things mood somewhat getting drunk mildly...

**Figure 2: Topics generated by LDA on mood transcripts**

## Models

- *TF-IDF:*
  *Term-frequency, inverse document frequency*

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

- *Latent Dirichlet allocation (LDA)*
  - *Unsupervised model used for topic modeling*
- *Logistic Regression (Max Entropy)*
  - Multinomial logistic regression
- *SVM with Gaussian kernel*
  - *Hyperparameters optimized with grid search*
- Multinomial NB utilized with TF-IDF

## Results:

### 3 Class: (Subclinical, Mild/Moderate, Moderately Severe/Severe)

| | Train Accuracy | Test Accuracy |
|---|---|---|
| LDA Mood LR (Multinomial) | 67.8 | 63.5 |
| **LDA Mood SVM** | **74.0** | **70.6** |
| User Data LR (Multinomial) | 59.9 | 57.1 |
| User Data SVM | 69.3 | 62.2 |
| TFIDF All Text Naive Bayes | 36.5 | 30.3 |
| TFIDF All Text LR | 60.9 | 55.2 |
| TFIDF All Text SVM | 65.8 | 65.4 |

Training set: 86 samples, / Test set: 37 samples

### 2 Class: Depressed/Not Depressed

| | Train Accuracy | Test Accuracy |
|---|---|---|
| LDA Mood LR | 79.9 | 77.0 |
| **LDA Mood SVM** | **77.9** | **80.3** |
| User Data LR | 75.9 | 71.9 |
| User Data SVM | 77.1 | 74.6 |
| TFIDF All Text Naive Bayes | 36.5 | 30.3 |
| TFIDF All Text LR | 77.6 | 72.3 |
| TFIDF All Text SVM | 73.2 | 73.5 |

Training set: 104 samples, / Test set:44 samples

## Conclusions/Future Work

- Topics generated by LDA useful for characterizing mood from text, which can be used to predict PHQ-9 outcome
- SVM model with LDA mood is the best model for two and three class, also most balanced accuracy per class
- Examine which features/topics are most predictive of depression level
- Use EmoLex, LIWC lexicons to generate more features

**References**

1. https://www.cs.jhu.edu/~mdredze/publications/2014_acl_mental_health.pdf
2. http://www.aclweb.org/anthology/W/W15/W15-12.pdf#page=13
3. https://www.aclweb.org/anthology/W/W14/W14-32.pdf#page=130
4. https://dl.acm.org/citation.cfm?id=2464480