Predicting outcomes in online chatbot-mediated therapy

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

Background

More than 50% of college students suggest symptoms of anxiety and depression in the previous year so severe that they couldn't function.[1] However, up to 75% of them do not access clinical services. While the reasons for this are varied, the ubiquity of free or inexpensive mental health services on college campuses suggest that service availability and cost are not primary barriers to care.[2] Like non-college populations, stigma is considered the primary barrier to accessing psychological health services.

One possible solution to the issue of stigma are online interventions where patients have been shown to be more comfortable in seeking help with semi-anonymity of the internet.[3] However, despite a few clinically successful programs, these have not been successful in user engagement and adherence with about 50% finishing each module potentially due to the loss of the human interactional quality.[4]

Woebot is a therapeutic chatbot on Facebook Messenger that was designed to take advantage of the lower barrier for stigma, scalability and 24/7 accessibility that advantage online interventions but also deliver its therapeutic content in a conversational, more human-like way to improve engagement and treatment adherence.

Since it was launched last summer by Alison Darcy, a former clinical psychology researcher at Stanford, Woebot has had millions of conversations with users online. Although it initially targeted younger college-age users, its user base has spread across all ages with an average age of 34.

Not only is Woebot based on evidence-based Cognitive Behavioral Therapy, but it's also the first therapeutic chatbot to have a study conducting at Stanford that demonstrates its greater efficacy in lowering symptoms of depression and anxiety versus a standard educational materials control.[5]

Project Objectives

Over the summer, I conducted a 400-person follow up study on the Woebot mental health chatbot to understand the relationship between user text input and standard psychological surveys that measure depression (PHQ-9), anxiety (GAD-7) and therapeutic alliance (WAI) at three timepoints: a baseline at enrollment, two weeks later and four weeks later. This research study is the main component of my senior thesis in Symbolic Systems and I am currently working on publishing a paper on the results with therapeutic alliance. My CS 229 project focuses on a different subset of my research. In this paper, I attempted to predict a user's severity of depression as defined by the PHQ-9 at the two week time points using both survey data and user transcript text data.

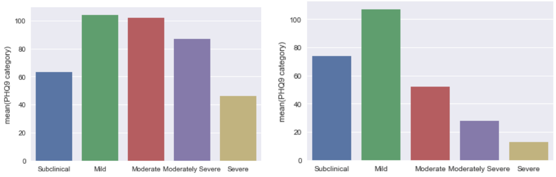


Figure 1: the distribution of PHQ-9 depression severity categories at baseline (top) and after two weeks (bottom). In order left to right: Subclinical, Mild, Moderate, Moderately Severe and Severe.

I utilized a two-class classifier, where users were classified as either Clinically Depressed (PHQ-9>5) or Subclinical (PHQ-9 <=5) , meaning they were not depressed by clinical guidelines. I also utilized a three-class classifier where users were classified into Not Clinically Depressed (PHQ9 <=5), Mild/Moderate Depression (10<= PHQ9 <20) and Moderately Severe/Severe Depression (20<= PHQ9 <27). These groupings of categories were chosen as they are commonly used to determine treatment options and course of treatment in clinical settings.

[1] Zivin et al, 2009
[2] Eisenberg et al, 2009
[3] Donkin, et al 2011
[4] Bickmore et al, 2005
[5] Fitzpatrick, Darcy, Vierhile, 2017.

The input to the classifier was (1) baseline survey data or (2) a TFIDF vector from all text data or (3) vector of the probability topics (from an LDA model) in given user's text mood data. Each of the three inputs were used to train two-class and three class logistic regression or SVM and in the case of (2) a Naive Bayes classifier.

Related Work

There were a handful of papers that I looked at language from online therapy sessions with a human therapist. [6] [7] [8] These papers provided a general framework for my work in suggesting daily moods mentioned to a human therapist are important as well as the therapeutic alliance score (in the survey data, 1) are required for effective improvement. A paper on text with a human therapist pointed out that lexicons have their limitations in working with short text messages and also suggested the power of LDA topic modelling to discover subtle patterns that reveal interesting connections between words in latent topics.[9]

I was most impressed by the technical approaches of papers analyzing massive datasets of mental health communities on Twitter and Reddit. In particular, I liked "Quantifying the Language of Schizophrenia in Social Media" where multiple models of text processing (including LDA) were used with SVM and Logistic Regression models to predict schizophrenia in users.

What is important to note is that a much more complex model using a Twitter dataset achieved a maximum of 74% accuracy in depression prediction.[10] Another highly cited study had a positive precision of 48% and negative of 68%.[11] This goes to show the challenges of depression prediction, even in a 2-class model. Depression, in particular, is difficult to classify due to the diversity of its presentation and comorbidity with other mental illnesses and overlapping symptoms with other mental illnesses.

Dataset and Features

My dataset came entirely from the study I ran in the summer with IRB approval. 274 users (68.5%) completed the study to the 2 week timepoint. Their deidentified survey and transcript data were used in this research.

To balance the training set by classes through random sampling, 86 users were used in the training set for the three class model. 19 and 18 users were used respectively for the validation and test sets. For the 2 class model, 104, 22, 22 users with balanced classes were used for the training, validation and test sets.

Survey Data (1)

Survey data included standard psychological measures such as the previously mentioned WAI, patient history (e.g. have you seen a therapist before?), demographics and user engagement data (daily usage percentage, words per day/message).

Categorical data points were processed to One Hot Encoding and missing data points were replaced with imputation to the average. All word count related variables were normalized by taking the log of their value and finally, all variables were normalized to have 0 mean and unit variance.

10 features out of 47 features were selected through a Recursive Feature search algorithm that utilized 5-fold Cross-Validation. I also utilized Random Forest to determine the relative importance of these features.

2) Text Data (M=274)

13,436 user messages to Woebot in the target 2 week span were used to generate this dataset.

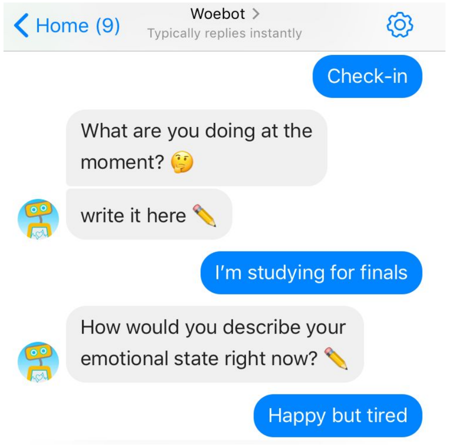
Stop words were removed using scikit-learn's standard list of stopwords. Having tried word counts, term frequency (tf) and term frequency inverse document frequency (tf-idf), I chose tf-idf as the vectorization of the text data. I tested out different feature set sizes between 100 and 6000 and found that performance did not vary significantly with larger feature set sizes..

[6] Van der Zanden, Rianne, et al., 2014.
[7] Svartvatten, et al 2015
[8] Dirkse, Dale, et al, 2015
[9] Howes, Purver, and McCabe. 2014
[10] De Choudhury, Munmun, et al. 2014
[11] Coppersmith, Dredze, and Harman, 2014.

Thus, I settled on using the 100 word features with the highest frequency.

3) LDA Topics Generated from Mood Data

The LDA topics were extracted from 24,634 user responses to the mood question in the Woebot daily check-in from all 400 users at timepoints outside the scope of this project. We did this to improve the quality of the topic modeling based on the assumption that mood topics for the 274 users in the first two weeks of this study were the same as mood topics of this study.



Check-in

What are you doing at the moment? 🤔

write it here ✏️

I'm studying for finals

How would you describe your emotional state right now? ✏️

Happy but tired

Figure 2: The daily Woebot user check-in. Mood text data was exclusively from the response to the second question which averaged 4.3 words per day.

Topic counts from 10 to 40 in increments of 5 were tried for the LDA model and 25 was chosen based on the internal consistency and coherence of the topics. We later tested out different topic counts and found 25 to be slightly better than the other topic counts, followed closely by 30 in model accuracy. For maximum accuracy, using all 25 topics were found to be optimal.

The LDA model was applied to 3948 user mood responses in the target 2 week span to generate a probability vector of topics for each user.

Methods

Weighted Term Frequency

For the input matrix for the LDA model, we used weighted tf (term frequency), weighted by the number of words in a message. For each word feature in the matrix:

$$weighted\ tf(word) = \frac{count\ of\ given\ word\ in\ message\ (tf)}{number\ of\ words\ in\ message}$$

TF-IDF

For the input matrix for the text data based classfier, we used tf-idf (term frequency), which is the raw word count per word feature weighted by the inverse document frequency. The inverse document frequency is the probability a word is in a document and measures how rare a given word is. For each word feature in the matrix:

$$tf - idf(word) = tf * (log(\frac{N}{|number\ of\ messages\ \ that\ contain\ word|}))^{-1}$$

where N is the total number of messages. This weighting allows for a unique words that appear in fewer messages to have higher scores and is often used in classification schemes.

Latent Dirichlet Allocation

LDA is an unsupervised generative learning model that assumes all words in a corpus belong to a number of latent distributions or "topics." LDA is identical to probabilistic latent semantic analysis (pLSA) except it uses a Dirichlet prior. Here is a rough sketch of how LDA works:

```
For every word w and topic t:
Calculate p(topic t | message m) and p(word w | topic t) that come from this word w.
    1. Assign w to a topic based on p(topic t | message m) * p(word w | topic t) which in LDA is assumed to be p(word w| topic t)
Repeat until convergence of words in topics
```

Naive Bayes

The Naive Bayes classifier assumes that all probabilities of words are independent and calculates the probability a given word is in a class P(word x | class). Then makes predictions based on using conditional probability with the assumption of independent probabilities with the following equation:

$$\hat{y} = \underset{k \in \{1,\dots,K\}}{\operatorname{argmax}} \ p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k).$$

Naive Bayes is known to work surprisingly well given its simplicity and provided a baseline for our models.

SVM with Gaussian Kernel

The Support Vector Machine is binary classifier that tries to find the hyperplane that separates classes with the maximum margin. We tried a Linear Kernel first but found that the Gaussian Kernel gave us better performance, as characterized below:

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right)$$

This kernel was used to minimize the following hinge loss function with regularization.

$$\left[\frac{1}{n}\sum_{i=1}^{n} \max\left(0, 1 - y_i(\vec{w} \cdot \vec{x}_i + b)\right)\right] + \lambda\|\vec{w}\|^2,$$

Logistic Regression

Logistic regression is a linear model classifer that minimizes the following loss function with cross-entropy loss where h(x) is the sigmoid function.

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} \mathrm{Cost}(h_\theta(x^{(i)}), y^{(i)})$$

$$\mathrm{Cost}(h_\theta(x), y) = -\log(h_\theta(x)) \qquad \text{if } y = 1$$

$$\mathrm{Cost}(h_\theta(x), y) = -\log(1 - h_\theta(x)) \qquad \text{if } y = 0$$

One vs One Multiclass Classifier

As logistic regression and SVMs are binary classifiers, the one vs. one method must be used for the three class model. Each classifier was trained on each pair of classes. To classify a given data point, the output of all the pair classifiers was added and the one with the largest score was outputted as the prediction value.

Experiments/Results/Discussion

Each model had hyperparameters tuned through grid search and 10 fold cross-validation was used to evaluate each model. The LDA mood model with SVM worked best with both three and two classes which is notable as it used a much smaller dataset even before dimension reduction through LDA than the all text model.

	Pre	Rec	Acc
User Data LR	39	58	58
User Data SVM	63	61	62.2
TFIDF All Text NB	13	35	30.3
TFIDF All Text LR	61	58	58.1
TFIDF All Text SVM	57	61	61.3
LDA Mood LR	66	65	64.5
LDA Mood SVM	**72**	**71**	**71.0**

Figure 3: Performance with 3 classes (Not Depressed, Mild/Moderate Depression, Moderately Severe/Severe)

	Pre	Rec	Acc
User Data LR	78	76	75.9
User Data SVM	80	78	77.1
TFIDF All Text NB	72	65	64.8
TFIDF All Text LR	70	94	72.9
TFIDF All Text SVM	76	76	75.6
LDA Mood LR	82	81	78.3
LDA Mood SVM	**82**	**81**	**81.1**

Figure 4: Performance with 2 classes (Not Depressed, Mild/Moderate Depression, Moderately Severe/Severe)

Accuracy: correct predictions / all predictions
Precision: true positives per class / (true positives + false positives) per class(averaged over all classes)
Precision: true positives per class / (true positives + false negatives) per class (averaged over all classes)

This suggests that simply mood data is a powerful enough predictor of depression class. Given that each mood user input averaged at 4.3 words per check-in, this amounts to about an average 60 words per user being predictive of depression class. Furthermore, the LDA mood models outperformed the User Data models suggesting that high quality, longitudinal, daily mood data is a stronger predictor than traditional psychological survey data.

The SVM model did not necessarily outperform logistic regression by much but did considerably better in precision with the the Mild/Moderate class in between the the other two classes. It is also our intuition that with more training data, the SVM model will improve in performance versus logistic regression as before hyperparameter tuning, it tended to achieve higher training set accuracy but lower validation set accuracy.

	Predicted Class			
Actual Class		0	1	2
	0 (Not depressed)	22	7.6	0.9
	1 (Mild/Moderate)	2.4	19.3	8.7
	2 (M. Severe/Severe)	0.9	6.6	23.6

Figure 5: Confusion matrix for the 3 class LDA SVM averaged over 10 fold cross-validation

	Predicted Class		
Actual Class		0	1
	0 (Not depressed)	40.2	16.3
	1 (Depressed)	10.2	44.3

Figure 6: Confusion matrix for the 2 class LDA SVM averaged over 10 fold cross-validation

Examining the confusion matrices for the 2 class, it seems the model was about equally precise for each class. For the 3 class model, it makes sense that the classifier had the most error classifying Mild/Moderate as Moderately Severe/Severe (and vice versa) and Not depressed as Mild/Moderate, given there are bound to users on class boundaries. However, I was pleasantly surprised with how well the three-class worked, achieving precision, recall and accuracy that approached that of the two-class.

LDA topics

We intentionally chose not to use the human-defined mood labels that Woebot uses internally for its data as we hypothesized that human-defined labels may not be the best way to characterize mood.

Topic #0: meh sure oke head nauseous ambivalent awful resentful
Topic #1: good pretty feel optimistic energetic stable trying engaged
Topic #2: exhausted accomplished sore lost high stress unsure thoughtful
Topic #5: stressed worried work day time
Topic #9: overwhelmed slightly things mood somewhat getting drunk mildly

Figure 7: Sample topics from the LDA

In the LDA results, we found topics that were what one would expect such as topic #1 that resemble human categories. However, we also found some surprising topics that did not immediately seem coherent but upon evaluation, made sense such as topic #2 where exhausted, accomplished and stress are in the same category. These fine-grained, nuanced categories are perhaps is a reason for the effectiveness of our model over the model trained with traditional psychological moods of users found in the survey data set.

Conclusion/Future Work

We found topics generated by LDA useful for characterizing mood from text, which in turn which can be used to predict PHQ-9 depression outcomes. For both two and three class models, SVM model with LDA mood had the highest overall accuracy and the best accuracy per class.

I am currently working on obtaining 10,000 more user transcripts to have more data to train topic models on for mood and for other user responses. It is my hope that adding additional features from Woebot modules besides mood will improve accuracy. I am also planning on examining which features/topics are most predictive of depression level which will me understand the data further and help guide future research.

Finally, I plan on trying to use the LIWC and EmoLex emotion lexicons to generate more features. However that may work out with this set, I remain confident that generating topic models specific to user interaction over text messages with Woebot will prove an effective strategy in improving my work.

Contributions

I would like to thank Professor Kara Fitzpatrick who was my UAR grant advisor over the summer and continues to be the principal investigator who supervises my clinical work. She has contributed deeply to my understanding of clinical psychology and conducting rigorous clinical research, which has made this paper possible. I would also like to thank my Symbolic Systems thesis advisor Professor Dan Jurafsky in CS/Linguistics who has provided incredible insights on understanding social language and invaluable direction in building these machine learning models.

References

Coppersmith, Glen, et al. "From ADHD to SAD: Analyzing the Language of Mental Health on Twitter through Self-Reported Diagnoses." *CLPsych@ HLT-NAACL* (2015): 1-10.

Coppersmith, Glen, Mark Dredze, and Craig Harman. "Quantifying mental health signals in Twitter." *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. 2014.

De Choudhury, Munmun, et al. "Predicting Depression via Social Media." *ICWSM* 13 (2013): 1-10.

Dirkse, Dale, et al. "Linguistic analysis of communication in therapist-assisted internet-delivered cognitive behavior therapy for generalized anxiety disorder." *Cognitive behaviour therapy* 44.1 (2015): 21-32.

Fitzpatrick, Kathleen Kara, Alison Darcy, and Molly Vierhile. "Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial." JMIR Mental Health 4.2 (2017): e19.

Howes, Christine, Matthew Purver, and Rose McCabe. "Linguistic indicators of severity and progress in online text-based therapy for depression." *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. 2014.

Pennebaker, J. W. (1993). Putting stress into words: Health, linguistic, and therapeutic implications. *Behaviour research and therapy*, *31*(6), 539-548.

Pennebaker, James W., Janice K. Kiecolt-Glaser, and Ronald Glaser. "Disclosure of traumas and immune function: health implications for psychotherapy." *Journal of consulting and clinical psychology* 56.2 (1988): 239.

Svartvatten, Natalie, et al. "A content analysis of client e-mails in guided internet-based cognitive behavior therapy for depression." *Internet Interventions* 2.2 (2015): 121-127.

Van der Zanden, Rianne, et al. "Web-based depression treatment: Associations of clients′ word use with adherence and outcome." *Journal of affective disorders* 160 (2014): 10-13.AP