Chatbot Development: Phase 1

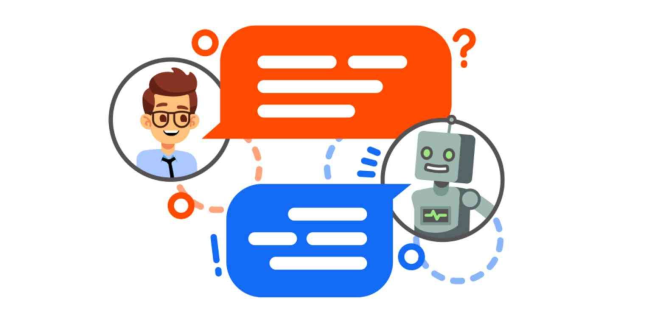
Model Predicting Reddit Posts in r/ADHD and r/OCD using NLP and Classification Modeling

ABC Research Group for the Department of Psychology, NUS
22 March 2021

Problem Statement

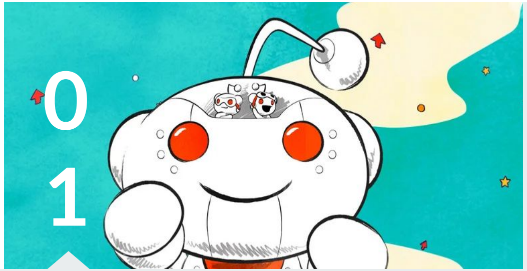
1. Volunteer manned chat- and phone lines are busier than ever due to the Covid-19 Pandemic.
2. Chatbot using NLP will be developed to reduce labour costs.
3. First step is to create a chatbot which can predict if an enquirer is talking about his/her ADHD or OCD!
4. Use Phase I chatbot to gather needed chat data for Phase II.

Can a chatbot determine if an enquirer is talking about ADHD or OCD?



Yes! If it is trained on data with the answers attached. We will train our chatbot with posts from Reddit.

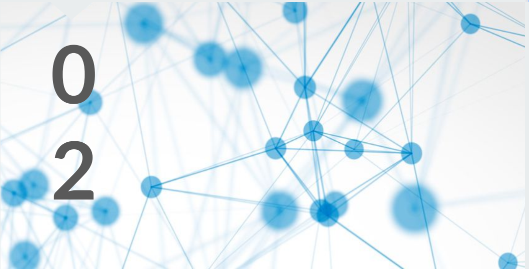
Methodology



01

Acquiring and Modifying Data

Reddit posts from r/ADHD and r/OCD were scraped and turned into fully text data.



02

Building the Model

A model which predicts posts membership to each subreddit was built through use of text data. It is validated with data not used to build model.

03

Studying Insights from Model

The models reveal insights which may be useful for further versions of the chatbot with more advanced features.

A look at the data

r/ADHD

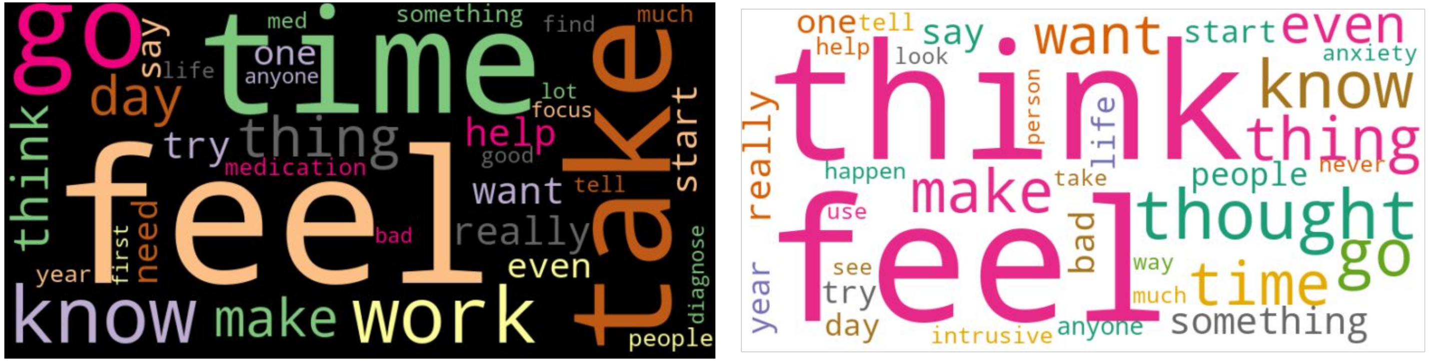
- 1.2 mil subscribers
- API yielded 800+ non-duplicated posts as data

r/OCD

- 112 K subscribers
- API yielded 800+ non-duplicated posts as data

A look at the data

Common Vocabulary



- Only main text and title of reddit posts were used.
- Very common vocabulary. Would be difficult to separate just using frequent words.
- Expected that fewer than 50% of words will be considered in model.

Modeling

Text is **Lemmatized.**

Vectorizers:

- Count
- TF-IDF

Classifiers:

- Logistic Regression
- Multinomial Naive Bayes

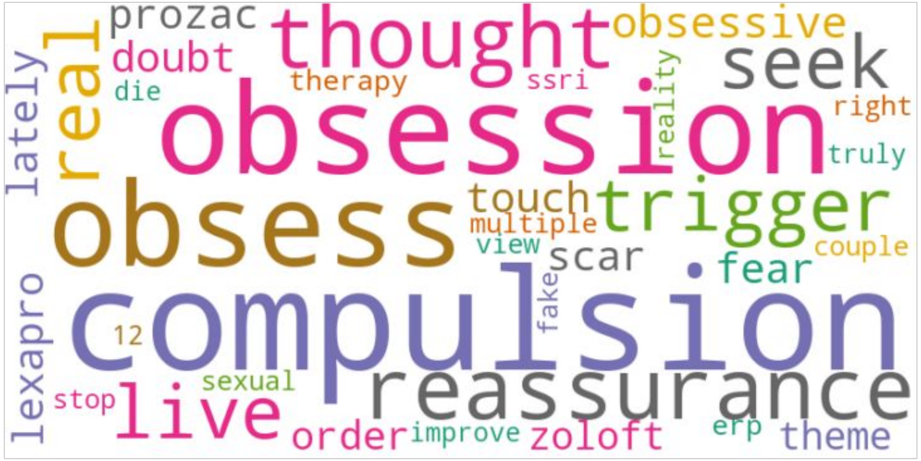
Results

	train_score	test_score	difference
gs3_Count_Lgr	0.866048	0.869464	-0.003416
gs2_TF_NB	0.899541	0.883450	0.016091
gs4_Count_NB	0.889391	0.864802	0.024589
gs_TF_Lgr	0.897194	0.871795	0.025399

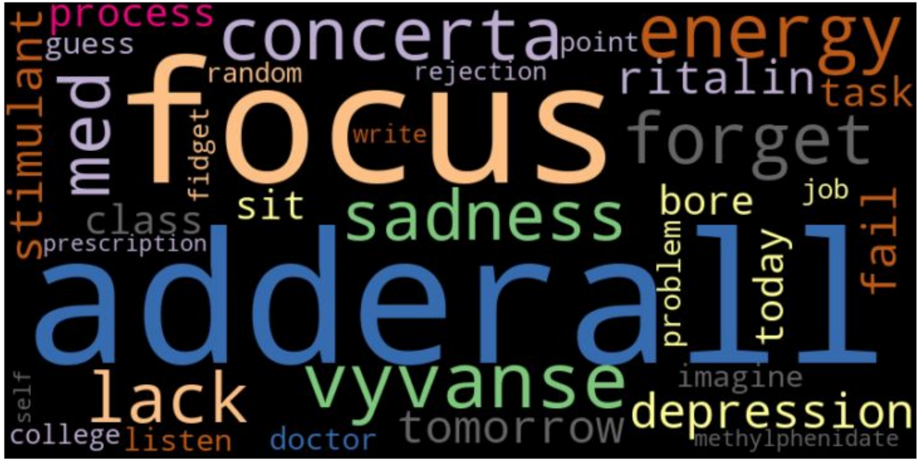
- TF-IDF Vectorizers had slightly better accuracy scores.

Important Words from Logistic Regression

r/OCD

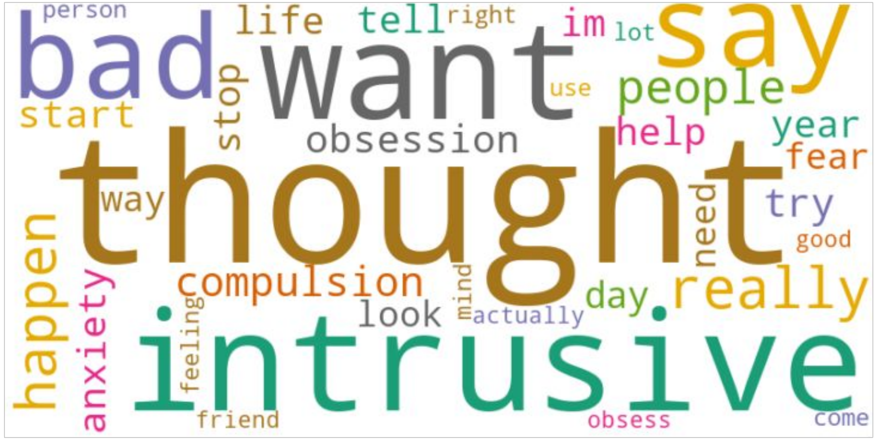


r/ADHD

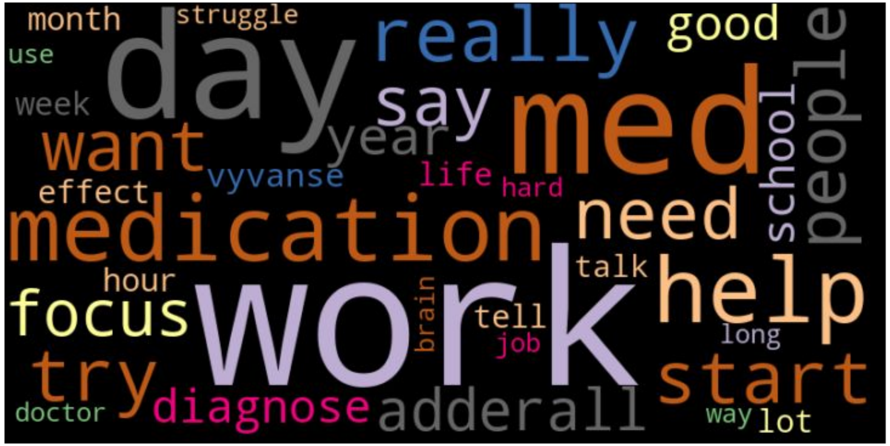


Important Words from M. Naive Bayes

r/OCD



r/ADHD



M. Naive Bayes is better for our chatbot

Logistic Regression cannot handle strong predictors for both classes. Good for prediction but may not be good for inference.

MNB can give insights useful for inference: r/ADHD tends to talk about activities and medications, r/OCD about their thoughts.

Conclusion

1. Chatbot Phase I will predict context between r/ADHD and r/OCD with ~90% accuracy.
2. M. Naive Bayes classifier more useful for gaining insight into new data as more data is created from the chatbot itself.
3. However, this is just the first step.

Future Work

1. Explore GloVe and Word2Vec algorithms for Phase II.
2. Gather conversation data from Phase I and retrain chatbot to predict in Singapore context.
3. Train model to handle more conditions than ADHD and OCD.

Thank you.
Q&A!