

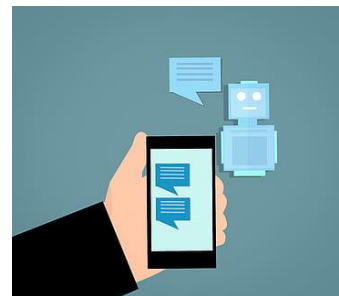
Classifying subreddits r/alcoholicsanonymous and r/stopsmoking

Providing rehabilitative services with digital
tools to better support recovering addicts



Content

- Introduction (Johnny)
- Data Cleaning (Saloni)
- Data Pre-Processing (June)
- Modelling (Matt)
- Model Evaluation (Guo Jun)
- Conclusions and Recommendations (Tze Ling)





Introduction

Background

- Who are we?
 - Scientist
 - Problem Solvers
 - Consultants



Introduction

Problem Statement

Problem: It can be hard to identify based on the subreddits, we aim to identify whether someone is reaching out to address a smoking addiction or alcoholism

Introduction

Objective

- To produce a personal assistant algorithm in a digital device that allows the author feel at ease when posting their personal problems.
- Personal device
- Rehabilitation centers
- Local community centers

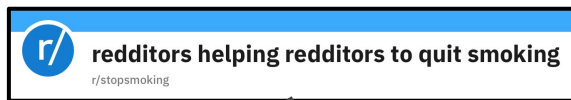


"And on my third day of sobriety...."



Data Cleaning

Web Scraping



r/stopsmoking = 0

+

r/alcoholicsanonymous = 1

2000 posts pulled from each subreddit

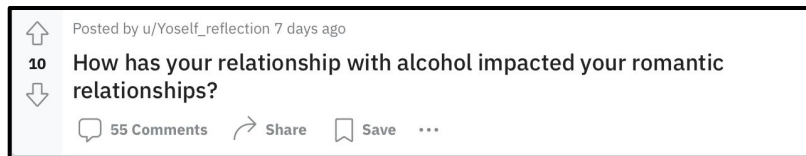
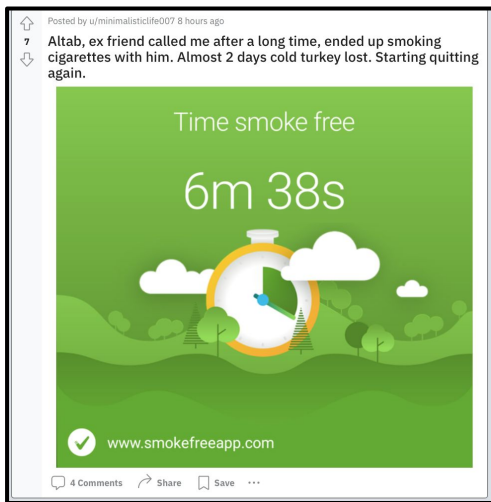
Assigned values

Merged datasets



Data Cleaning

Filtering posts





Data Cleaning

Dropping unwanted columns and characters

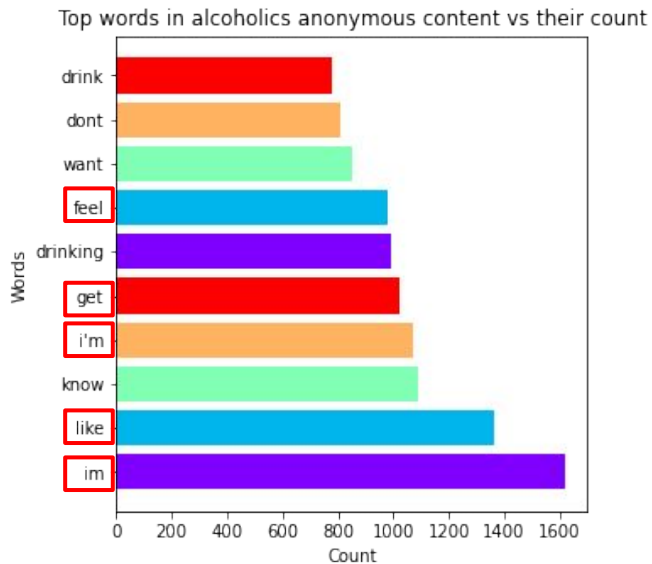




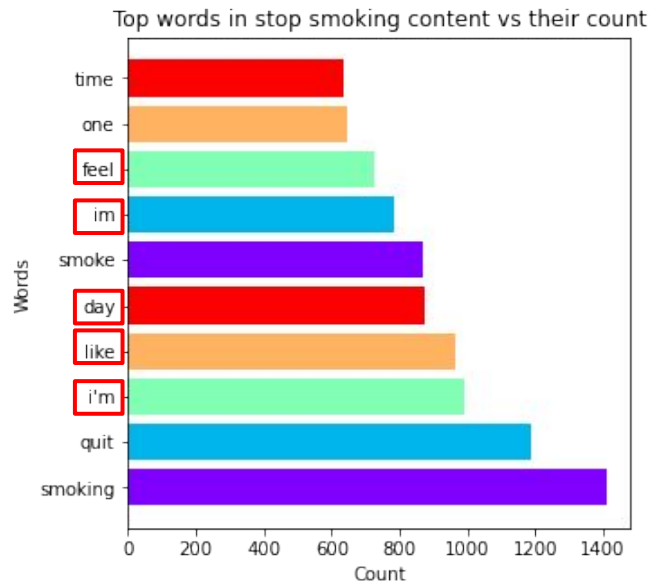


Data Cleaning

Preliminary EDA



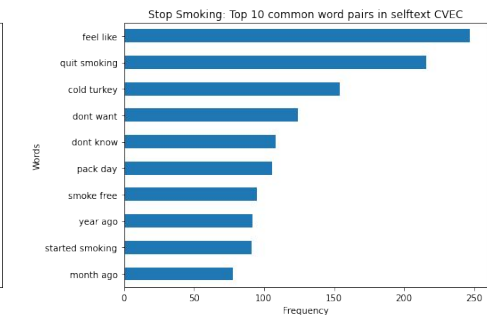
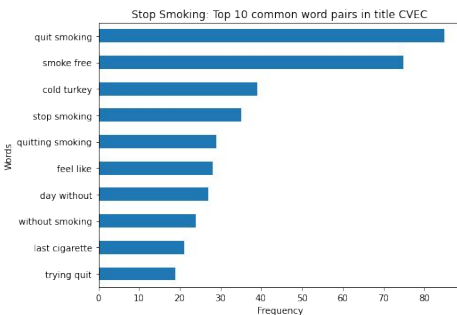
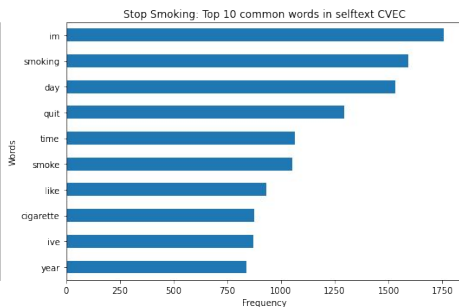
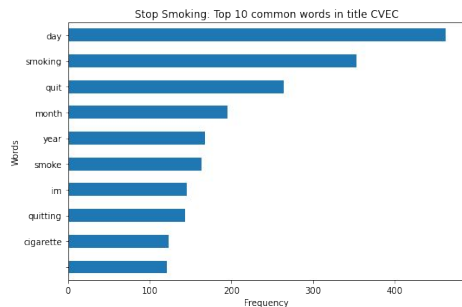
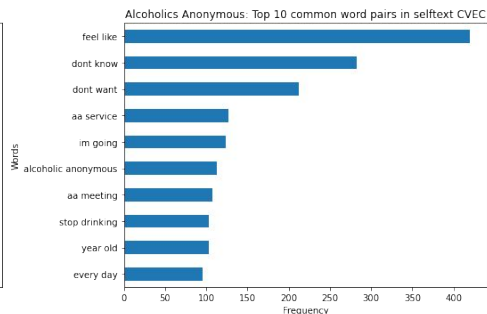
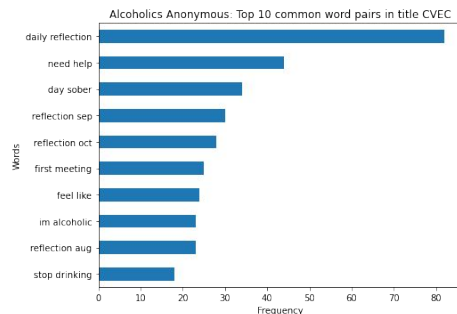
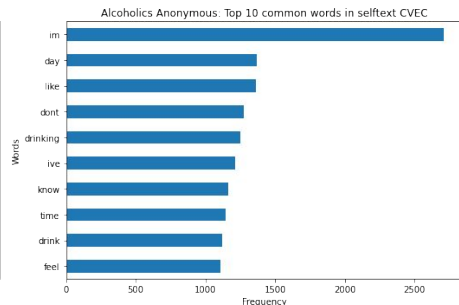
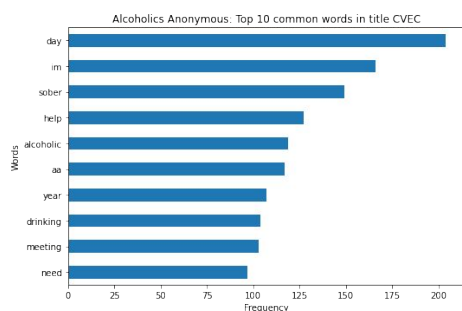
r/alcoholicsanonymous



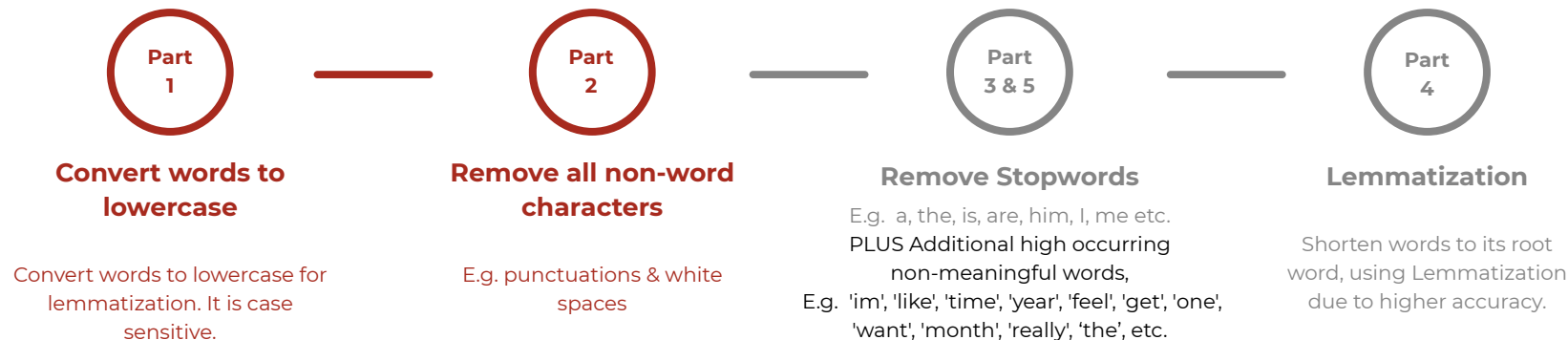
r/stopsmoking



EDA Barcharts after processing(single words and word pairs) - this is before custom stop words were removed



Data Pre-processing



Part 1

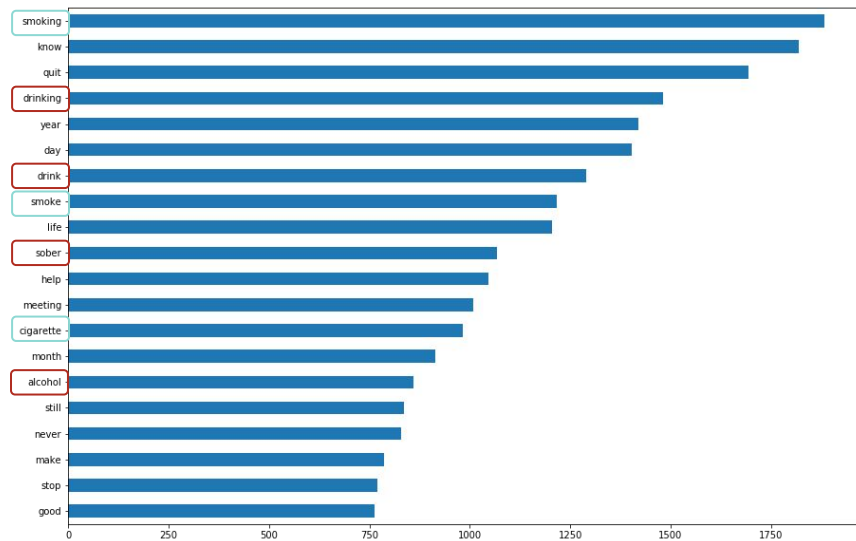
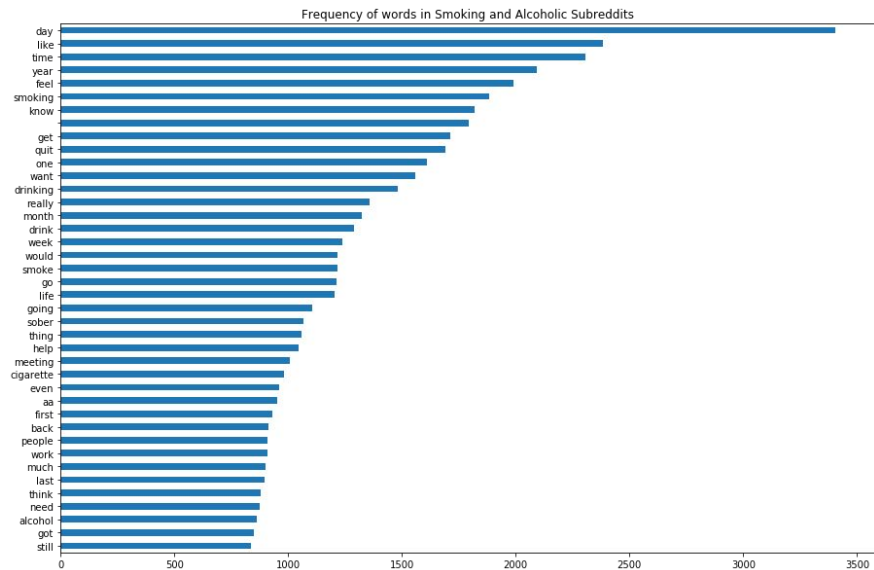
```
In [48]: 1 df1.loc[3]['text']
Out[48]: "Rehab and working from home??? So, does anyone know if you can go to rehab and still keep your job if you work from home?\n\nI need a retreat but I can't lose my job. Need to get off drugs and alcohol but also need my job.\n\nCan any one help please? I'm based in the UK."
```

Part 2, 3, 4

```
In [71]: 1 df.loc[3]['text']
Out[71]: 'rehab and working from home??? so, does anyone know if you can go to rehab and still keep your job if you work from home? i need a retreat but i cant lose my job. need to get off drugs and alcohol but also need my job. can anyone help please? im based in the uk.'
```

```
In [72]: 1 df.loc[3]['text_lemm']
Out[72]: 'rehab working home anyone know go rehab still keep job work home need retreat cant lose job need get drug alcohol al so need job anyone help please based uk'
```

EDA (Additional Stopwords)



Top words are more helpful in terms of classification

EDA (Additional Stopwords)

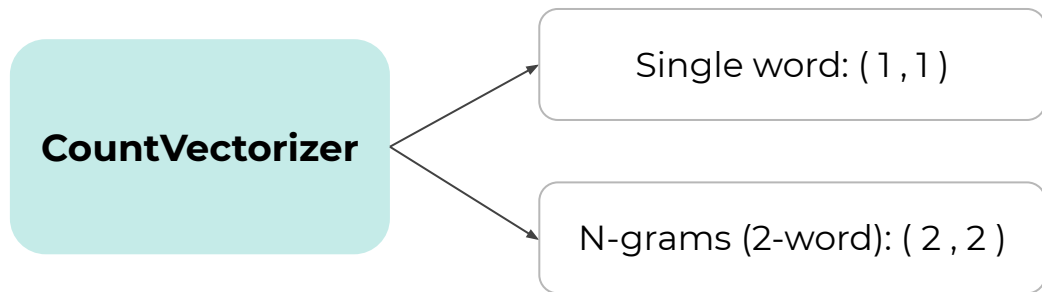


r/ Alcoholics Anonymous



r/ Stop Smoking

Vectorizers for modeling



TF-IDF
(Term Frequency -
Inverse Document
Frequency)

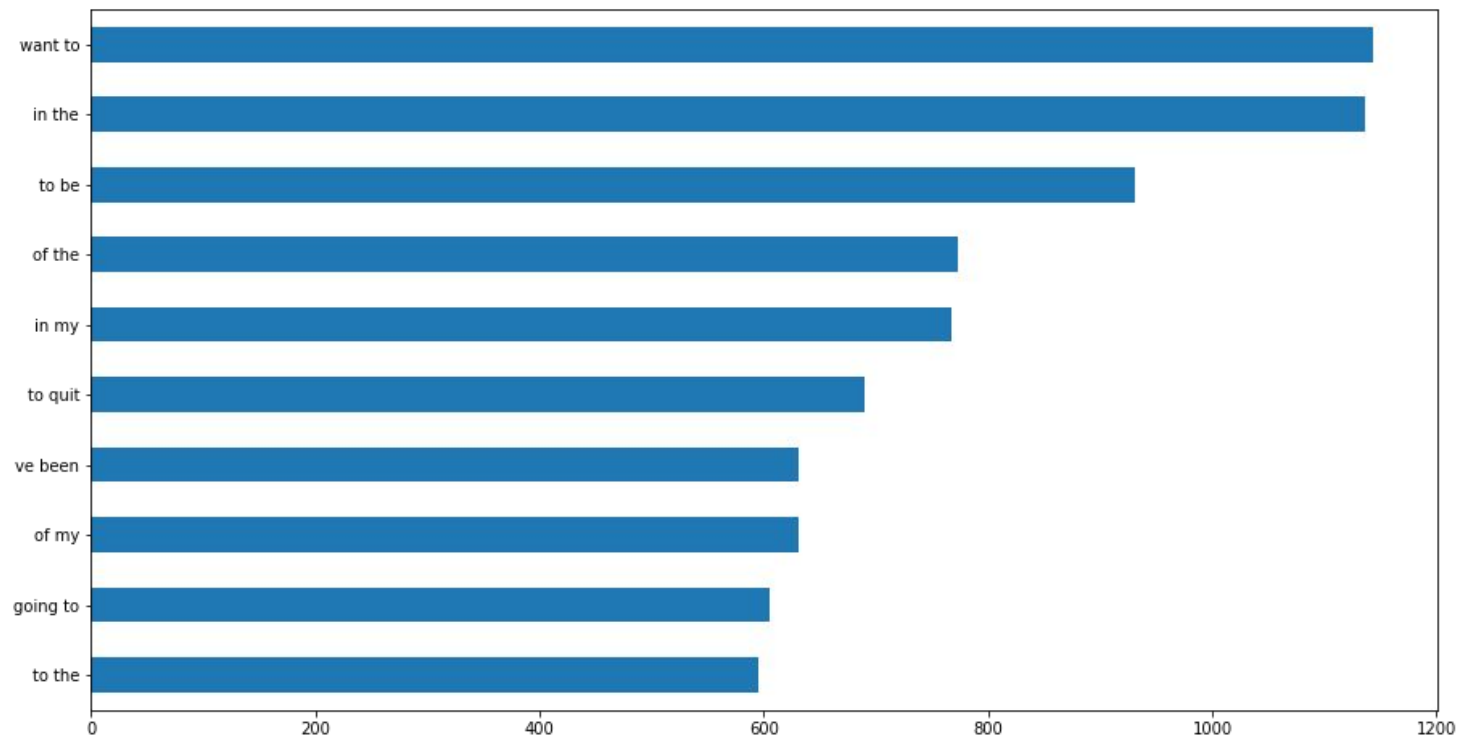
Bernoulli Naive Bayes

Multinomial Naive Bayes

Gaussian Naive Bayes

K-Nearest Neighbour

N-grams CountVectorizer (2 , 2)





Machine Learning Models



Baseline Model

Choosing Majority Words in Data

Looks at percentage of data in smoking or alcoholism

Decides the entire dataset by using the majority proportion in data



Our Model

Naive Bayes Model:

- Bernoulli
- Multinomial
- Gaussian

K-Nearest Neighbour Classifier



Why Naive Bayes Classifier

Works quickly and save time

Assumes independence of feature words

More suitable for our categorical variable from our count vectorized data

Suitable for multi-class prediction



K-Nearest Neighbour Classifier

Makes no assumption about our data distribution

Can be used in multi-class classification problems

Deals with outliers much more easily

Easier to implement



Disadvantages

Naive Bayes assumes all features are independent (Not true for most cases)

Naive Bayes cannot take into account additional features absent in its training dataset

KNN Model can be computationally expensive (Especially on memory)

KNN Model is highly dependent on the quality of our data

KNN Model may be slow in prediction when given large data



Comparison to current model

Our model has a much higher degree of accuracy (95% up from 55%)

Much less misclassifications (5% down from 45%)

Takes words into account for classification

Not computationally expensive (Naive Bayes works fast)

Can be integrated into web page frontends



Model Evaluation

Ngram (1,1)	Naive Bayes				KNN Model
	Bernoulli	Multinomial	Gaussian	Optimised	
Test	0.85	0.957	0.771	0.918	0.907
Train	0.884	0.985	0.942	0.950	0.862
ROC AUC	0.93	0.99	0.78	0.77	0.96



Model Evaluation

Ngram (2,2)	Naive Bayes				KNN Model
	Bernoulli	Multinomial	Gaussian	Optimised	
Test	0.587	0.928	0.893	0.939	0.785
Train	0.556	0.999	1.0	0.994	0.565
ROC AUC	0.76	0.98	0.89	0.89	0.57



Model Evaluation

TF-IDF	Naïve Bayes				KNN Model
	Bernoulli	Multinomial	Gaussian	Optimised	
Test	0.586	0.954	0.760	0.945	0.958
Train	0.549	0.988	0.951	0.988	0.943
ROC AUC	0.5	0.99	0.77	0.77	0.99



Best Model

(1,1) Multinomial Naive Bayes

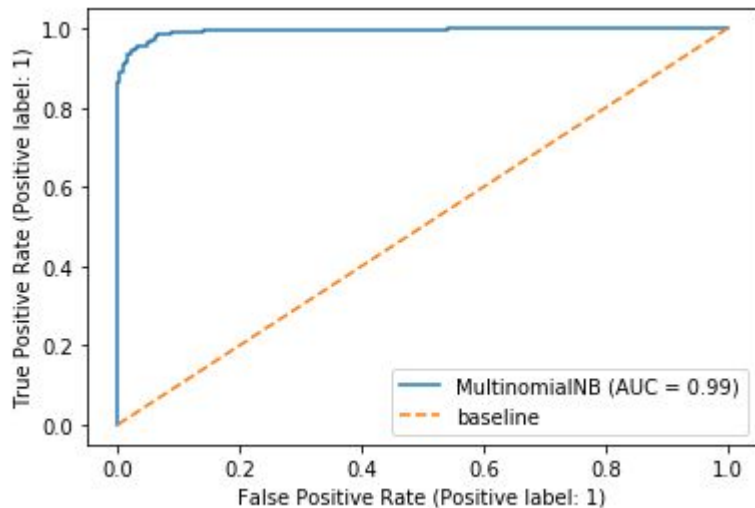
	precision	recall	f1-score	support
0	0.93	0.97	0.95	273
1	0.98	0.95	0.96	387
accuracy			0.96	660
macro avg	0.95	0.96	0.96	660
weighted avg	0.96	0.96	0.96	660

TF-IDF Multinomial Naive Bayes

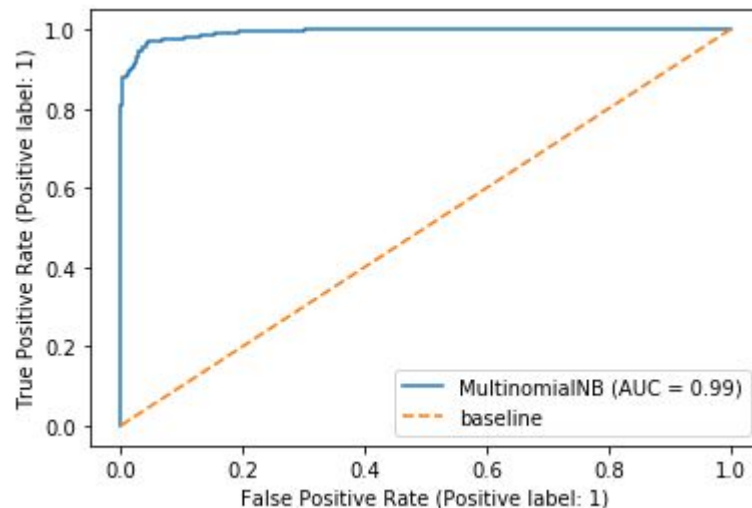
	precision	recall	f1-score	support
0	0.96	0.93	0.94	273
1	0.95	0.97	0.96	387
accuracy			0.95	660
macro avg	0.96	0.95	0.95	660
weighted avg	0.95	0.95	0.95	660



Multinomial Naive Bayes



Ngram (1,1)



TF-IDF



Feature Importance



Posted by u/NonnyMouse69 2725 days 4 years ago

For the Newcomers: "IWNDWYT"

IWNDWYT. You may see it multiple times as you surf [r/stopdrinking](#).

I

Will

Not

Drink

With

You

Today

This is a uniquely [r/stopdrinking](#) thing that we use as a sign of solidarity. Today I will stay sober, and so will you..it is a statement of support, and a recognition that no one does this alone.

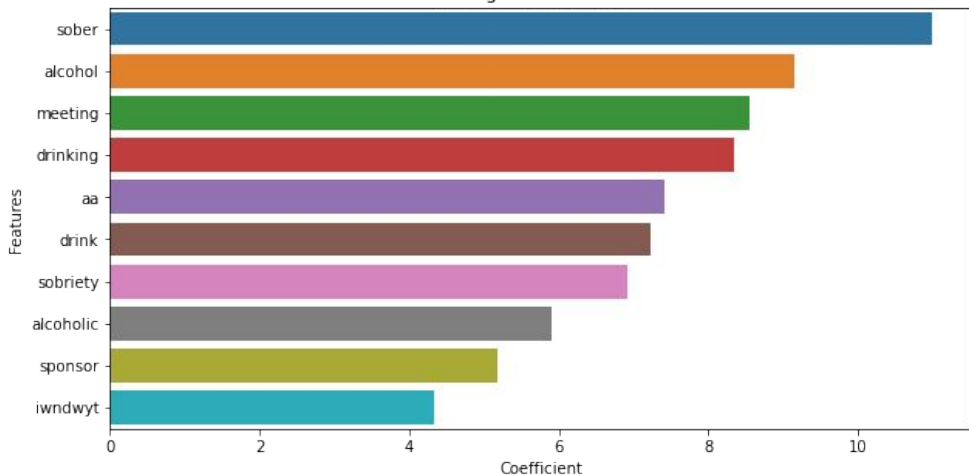
Stay sober my friends, IWNDWYT

Mouse

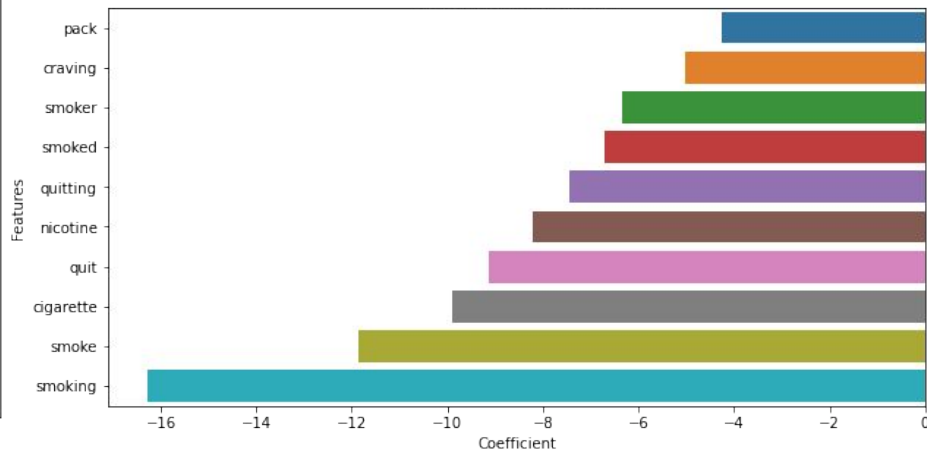
12 Comments Award Share Save Hide Report

94% Upvoted

Ten highest coefficients



Ten lowest coefficients





Conclusions & Recommendations

Chatbot/Online Platform

- Use the algorithm to determine target words for classification
- Based on the classification, identify the treatment need of the person and provide relevant recommendations



Conclusions & Recommendations

Hi, I want to get sober.

Please find more resources at
<link for alcoholism resources>
while we get a medical
professional to assist you.



Conclusions & Recommendations

I need help

How many alcoholic beverages
do you consume in a week?

Do you smoke?



Next Steps

Stopwords - further identify customized stopwords to improve the predictive power of the model

Apply larger N-gram vectorization to see if it yields more usable data

Expand the scope of data collection - posts may contain related issues, not limited to alcoholism and smoking



Thank
You