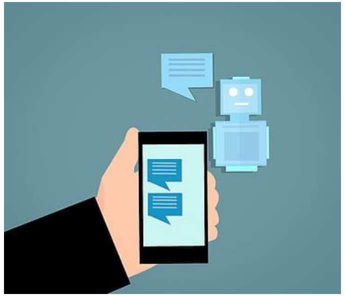
Classifying subreddits r/alcoholicsanonymous and r/stopsmoking

Providing rehabilitative services with digital tools to better support recovering addicts

Team 4: Johnny, Saloni, June, Matt, Guo Jun, Tze Ling

Content

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



Introduction

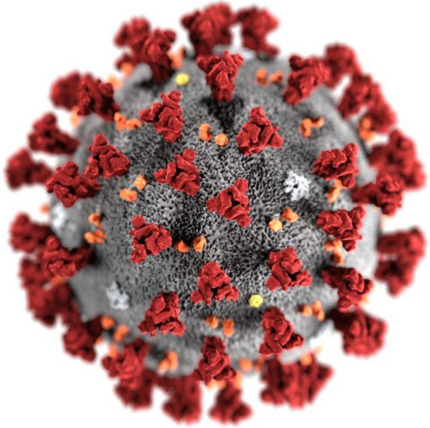
Background

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

Introduction

Problem Statement

- COVID-19
- Alcoholics
- Heavy Smokers



Introduction

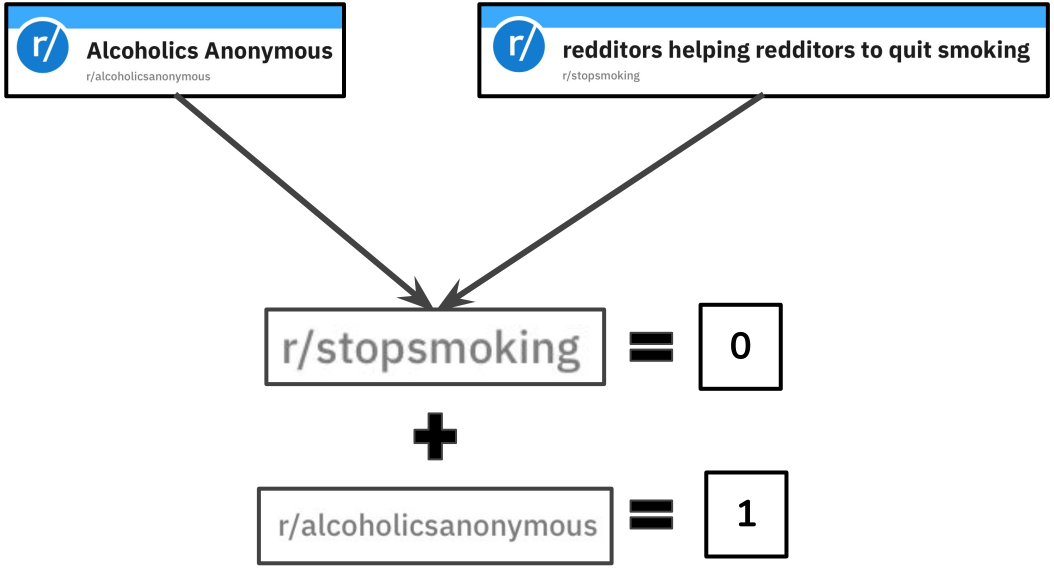
Objective

- To devise a machine learning algorithm
- To apply our tool into digital device(s)
- Share info to proper centers/facilities
- Rehabilitation centers
- Local community centers

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

Data Cleaning

Web Scraping



Alcoholics Anonymous

r/alcoholicsanonymous

redditors helping redditors to quit smoking

r/stopsmoking

r/stopsmoking = 0

+

r/alcoholicsanonymous = 1

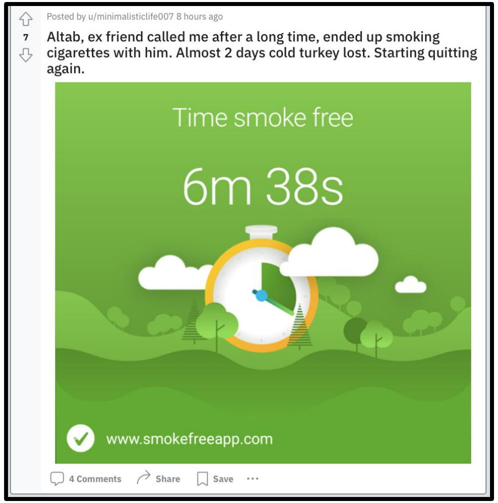
2000 posts pulled from each subreddit

Assigned values

Merged datasets

Data Cleaning

Filtering posts

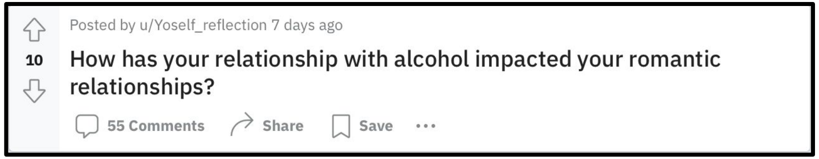


Posted by u/minimalisticlife007 8 hours ago

7

Altab, ex friend called me after a long time, ended up smoking cigarettes with him. Almost 2 days cold turkey lost. Starting quitting again.

✗



Posted by u/Yoself_reflection 7 days ago

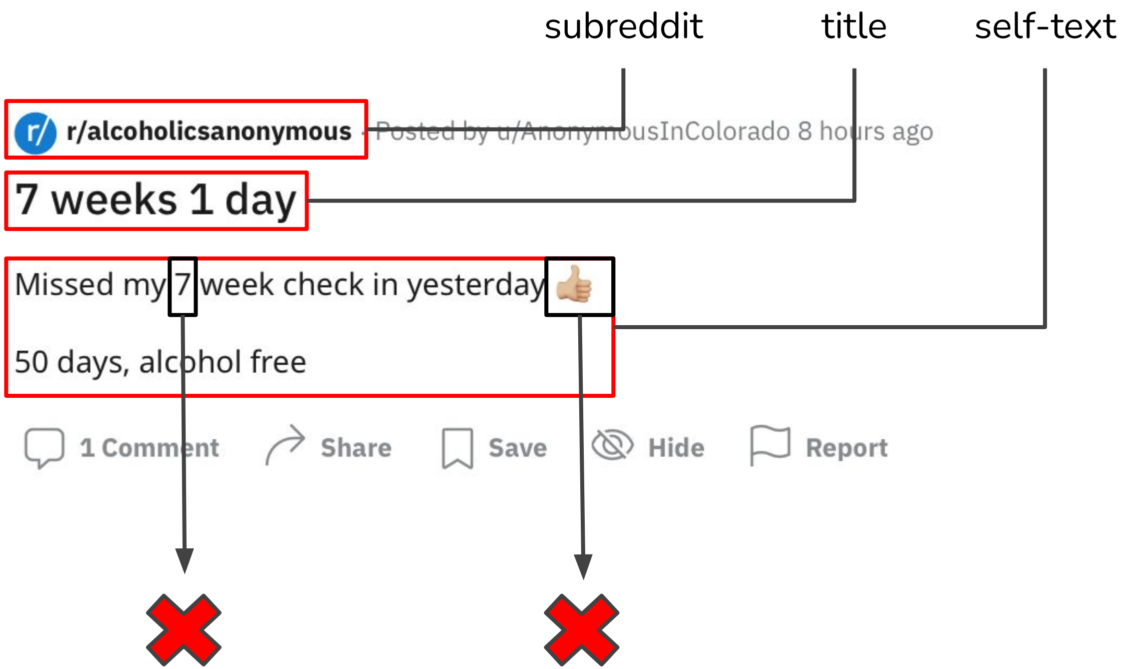
10

How has your relationship with alcohol impacted your romantic relationships?

✗

Data Cleaning

Dropping unwanted columns and characters



r/alcoholicsanonymous · Posted by u/AnonymousInColorado 8 hours ago

7 weeks 1 day

Missed my 7 week check in yesterday 👍

50 days, alcohol free

Data Cleaning

Preliminary EDA

would
drinking
sober
alcohol
one
get
work
like
drink
time
im
meeting
day
want
know
i'm
life
feel
people
thing
really
feeling
year
going
alcoholic

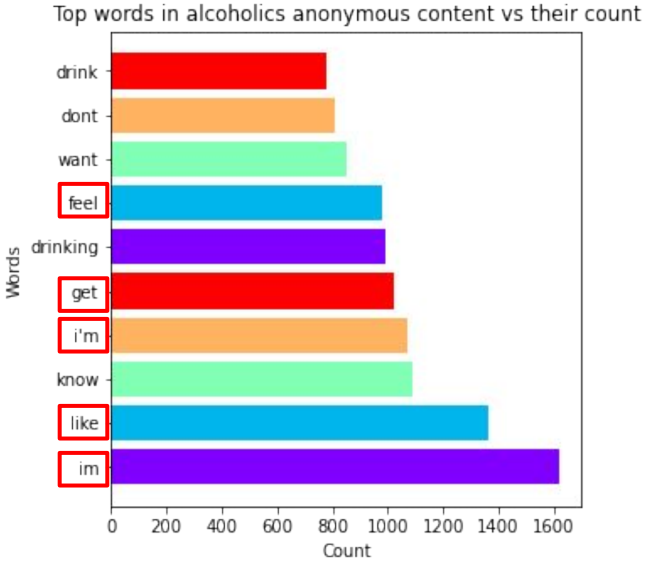
r/alcoholicsanonymous

smoking
nicotine
still
would
much
people
year
get
feel
think
never
smoked
week
quit
day
time
i'm
one
month
quitting
cigarette
smoker
craving
really
im
feeling
also
since
going
thing
like
life
help
good
pack
want
smoke
know
today
trying

r/stopsmoking

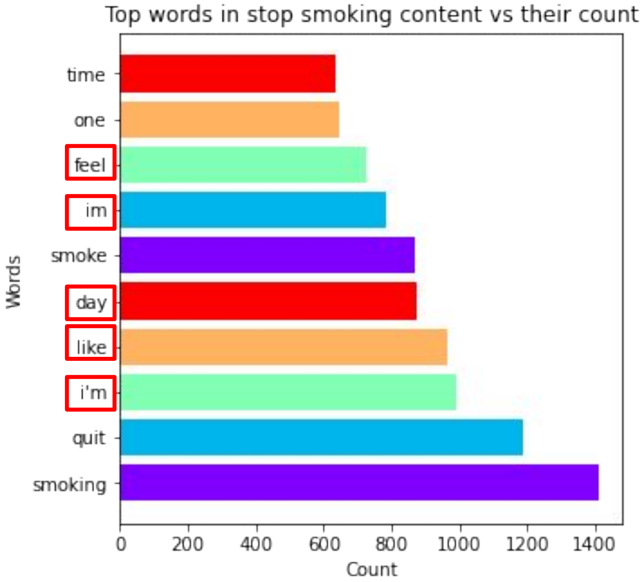
Data Cleaning

Preliminary EDA



Top words in alcoholics anonymous content vs their count

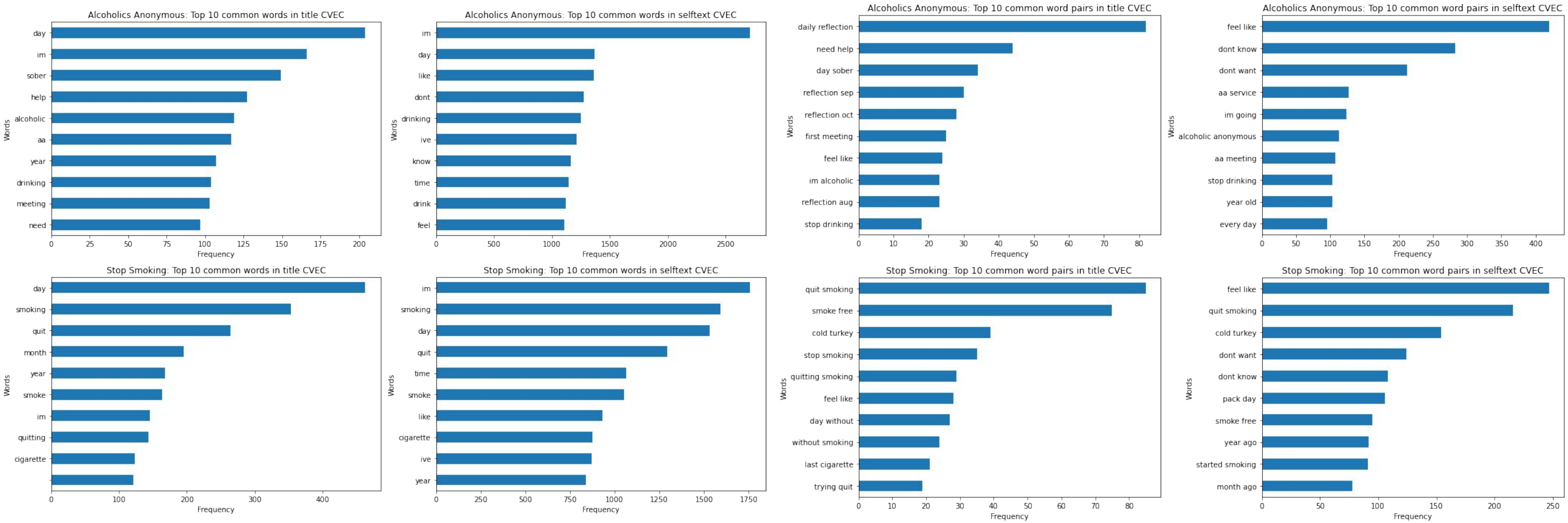
r/alcoholicsanonymous



Top words in stop smoking content vs their count

r/stopsmoking

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



Data Pre-processing

Part
1
Part
2
Part
3 & 5
Part
4

Convert words to lowercase

Convert words to lowercase for lemmatization. It is case sensitive.

Remove all non-word characters

E.g. punctuations & white spaces

Remove Stopwords

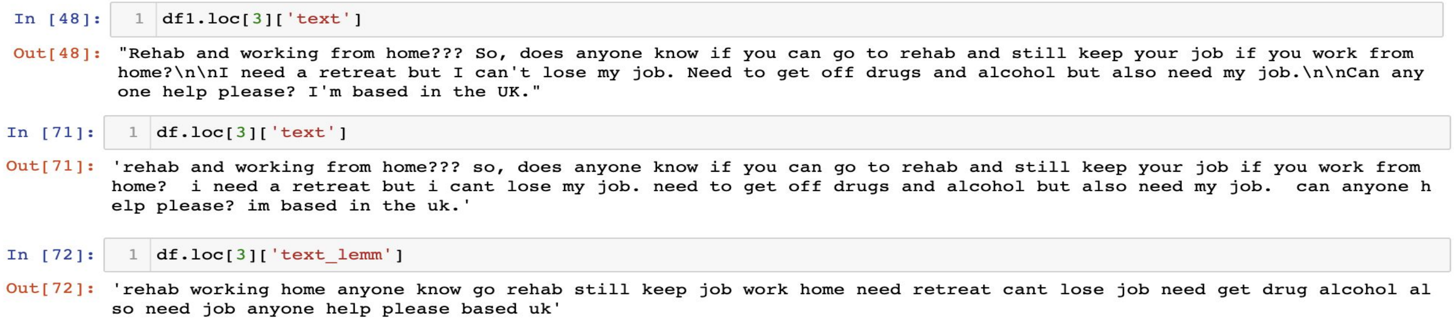
E.g. a, the, is, are, him, I, me etc.
PLUS Additional high occurring non-meaningful words,
E.g. 'im', 'like', 'time', 'year', 'feel', 'get', 'one', 'want', 'month', 'really', ‘the’, etc.

Lemmatization

Shorten words to its root word, using Lemmatization due to higher accuracy.

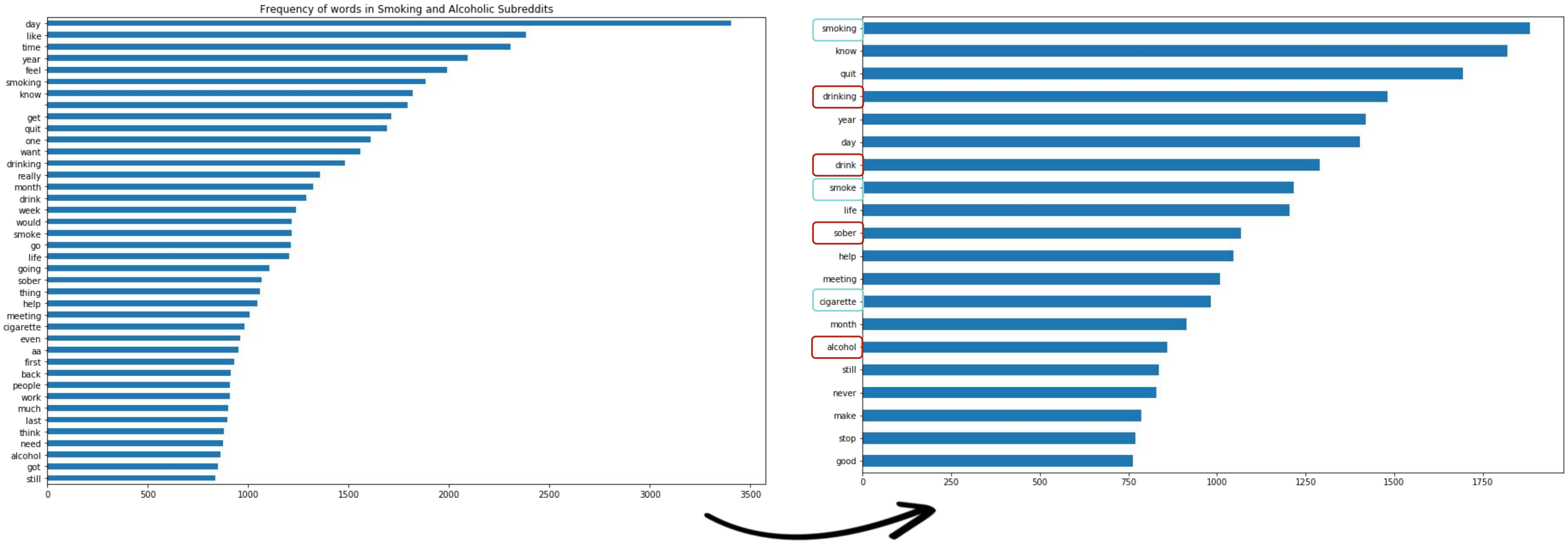
Part 1

Part 2, 3, 4



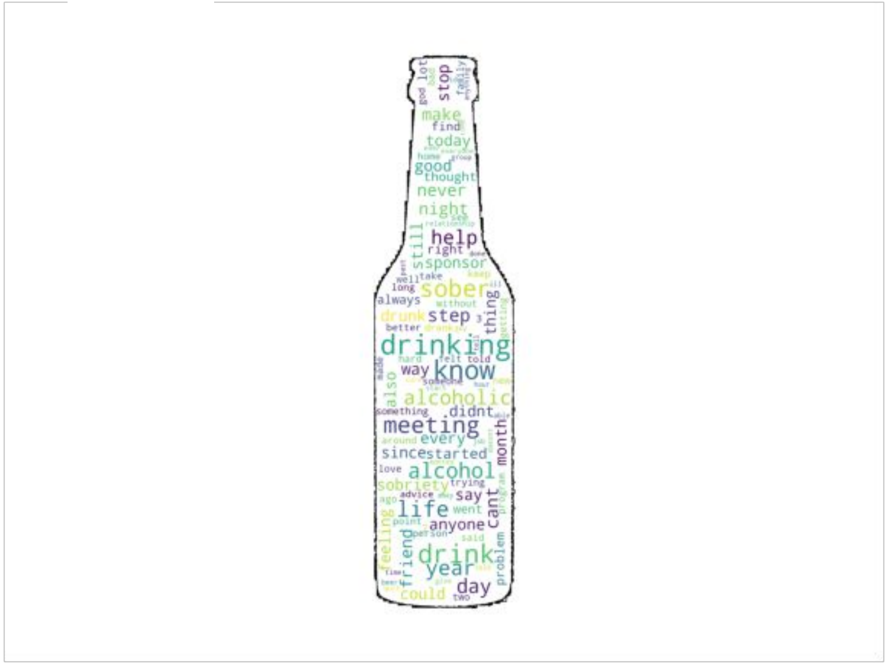
```
In [48]:  df1.loc[3]['text']

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."

In [71]:  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 h
          elp please? im based in the uk.'

In [72]:  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)

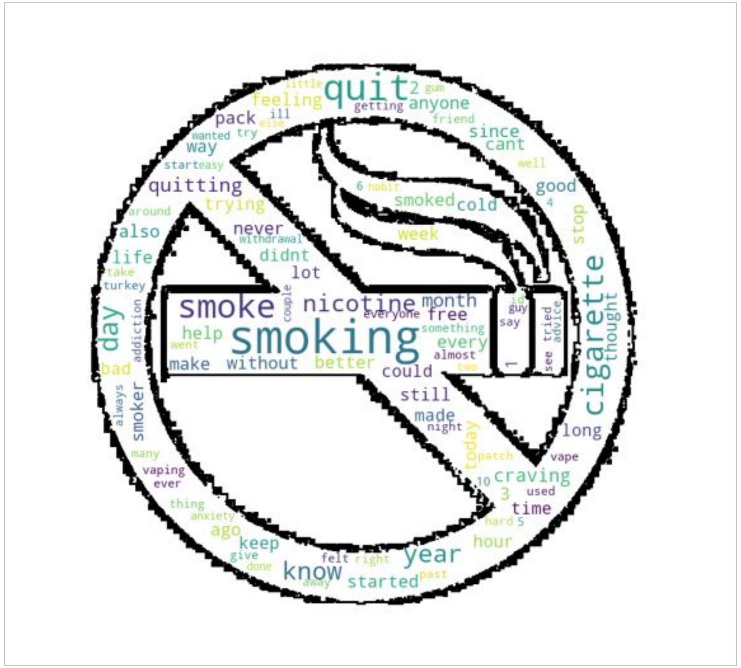


Top words are more helpful in terms of classification

EDA (Additional Stopwords)

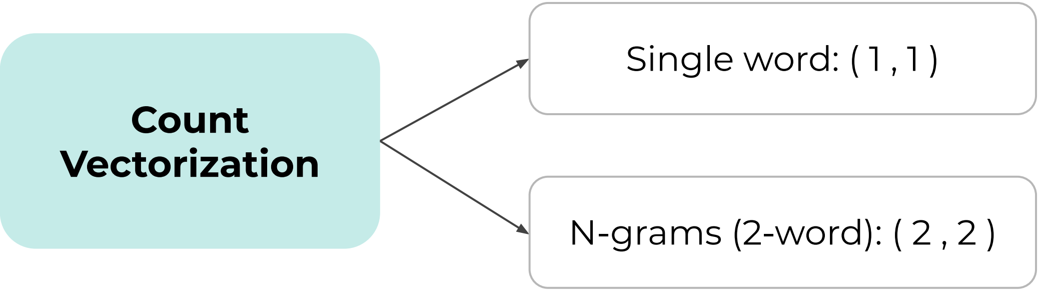


r/ Alcoholics Anonymous



r/ Stop Smoking

Vectorizers for modeling



Count Vectorization

Single word: (1 , 1)

N-grams (2-word): (2 , 2)

TF-IDF Vectorization
(Term Frequency - Inverse Document Frequency)

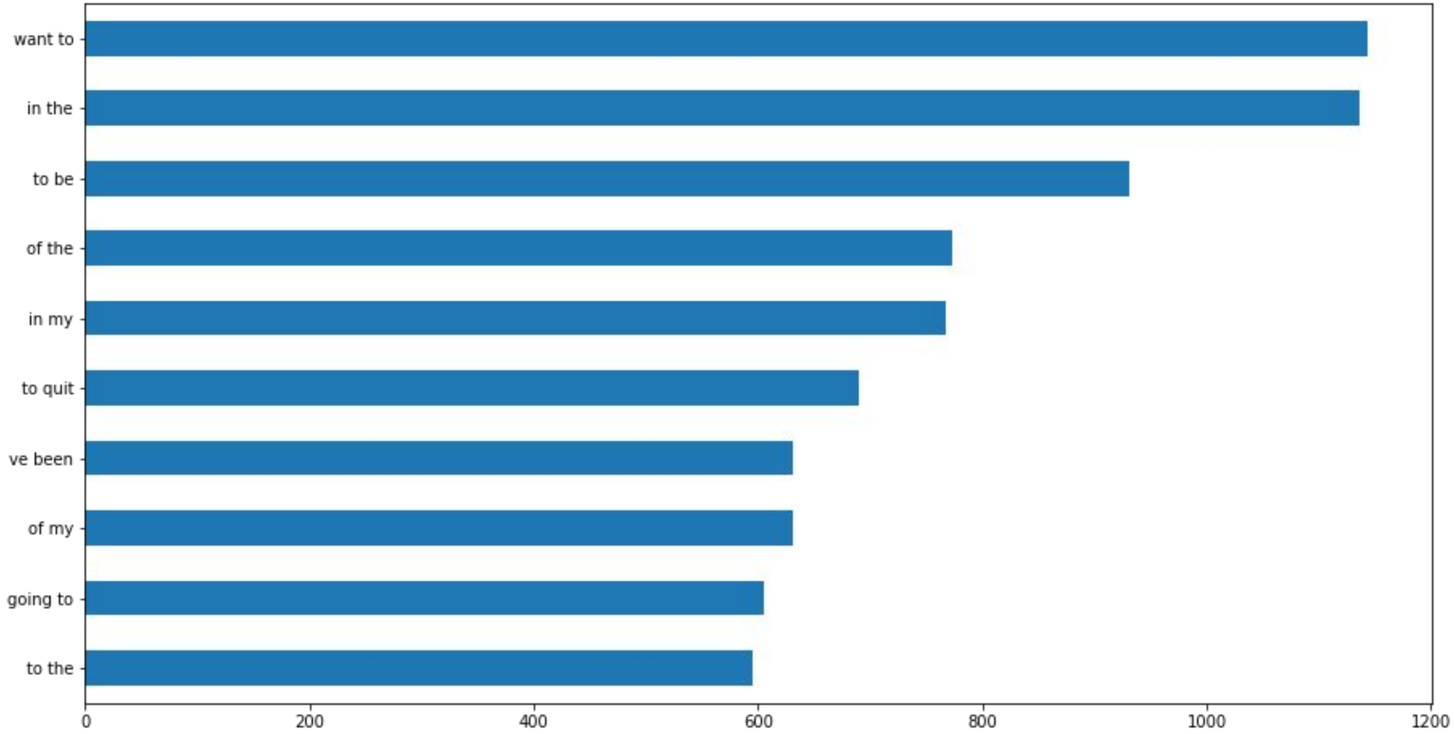
Bernoulli Naive Bayes

Multinomial Naive Bayes

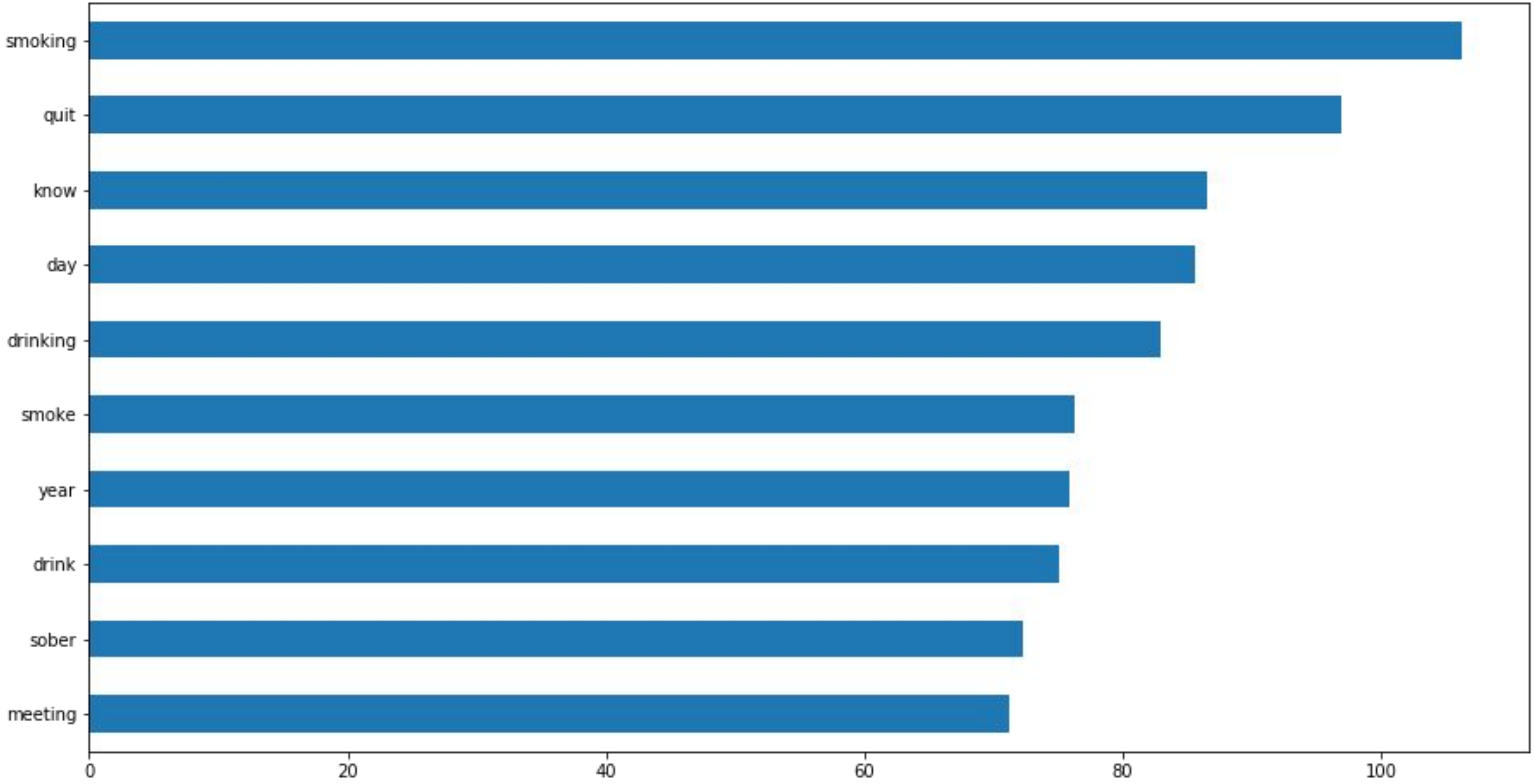
Gaussian Naive Bayes

K-Nearest Neighbour

N-grams CountVectorizer (2 , 2)



TF-IDF Vectorizer



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 it’s 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**

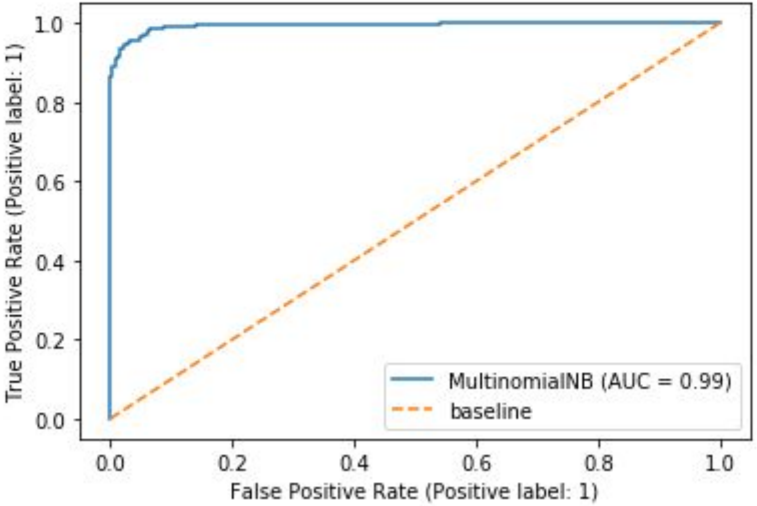
Model Evaluation

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

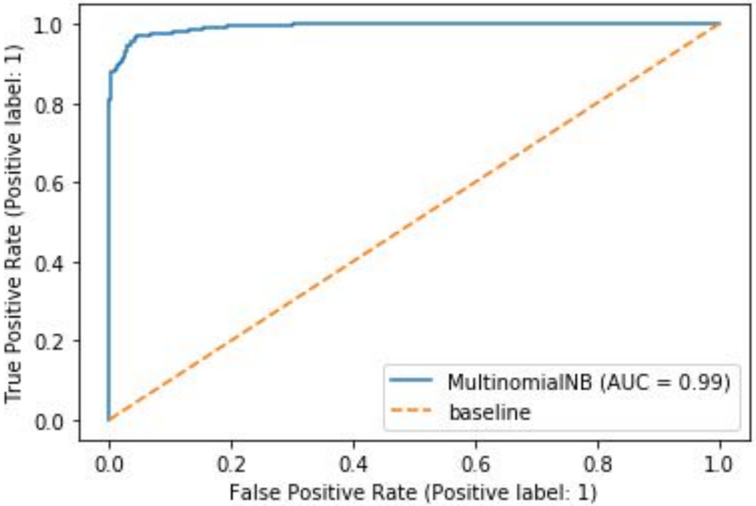
Model Evaluation

TF-IDF	Naive 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**

Multinomial Naive Bayes



Ngram (1,1)



TF-IDF

Best Model

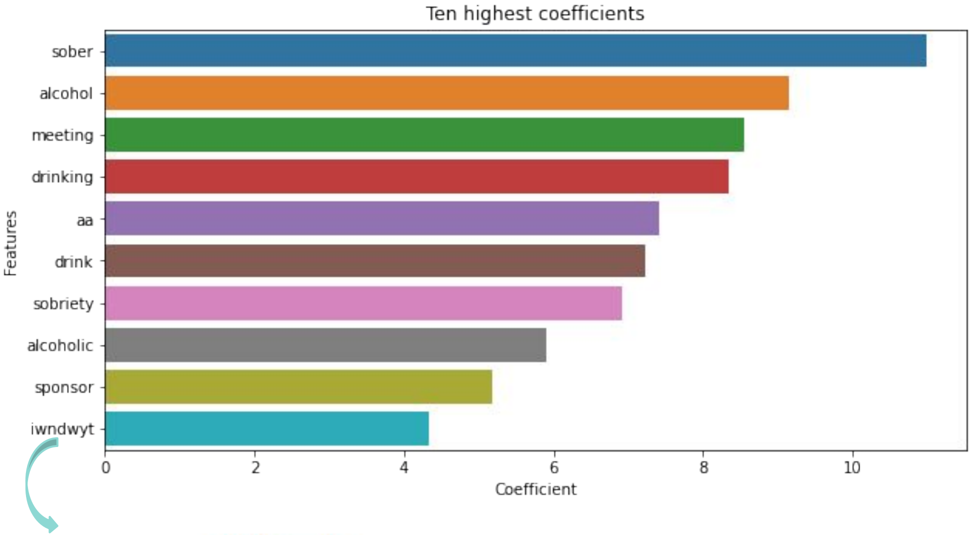
(1,1) Multinomial Naive Bayes

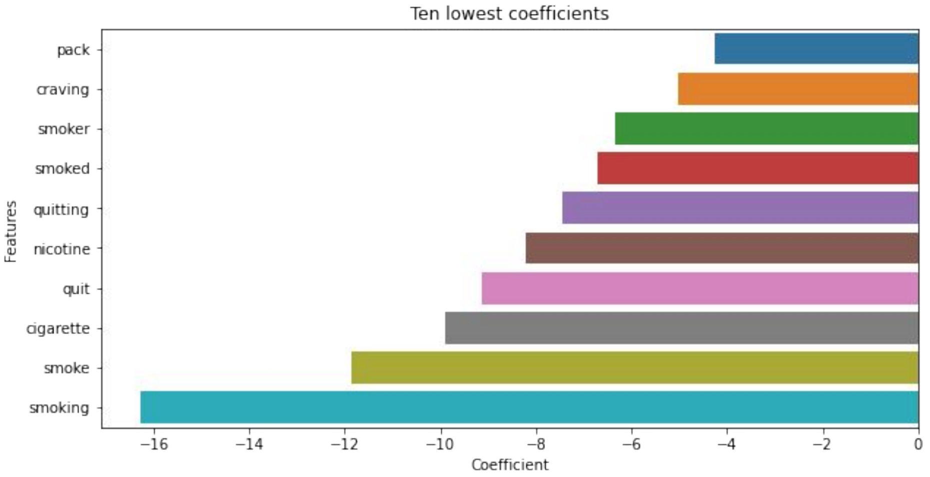
```
              precision    recall  f1-score   support

           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
```

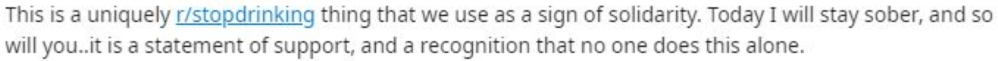
TF-IDF Multinomial Naive Bayes

```
              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
```

Feature Importance







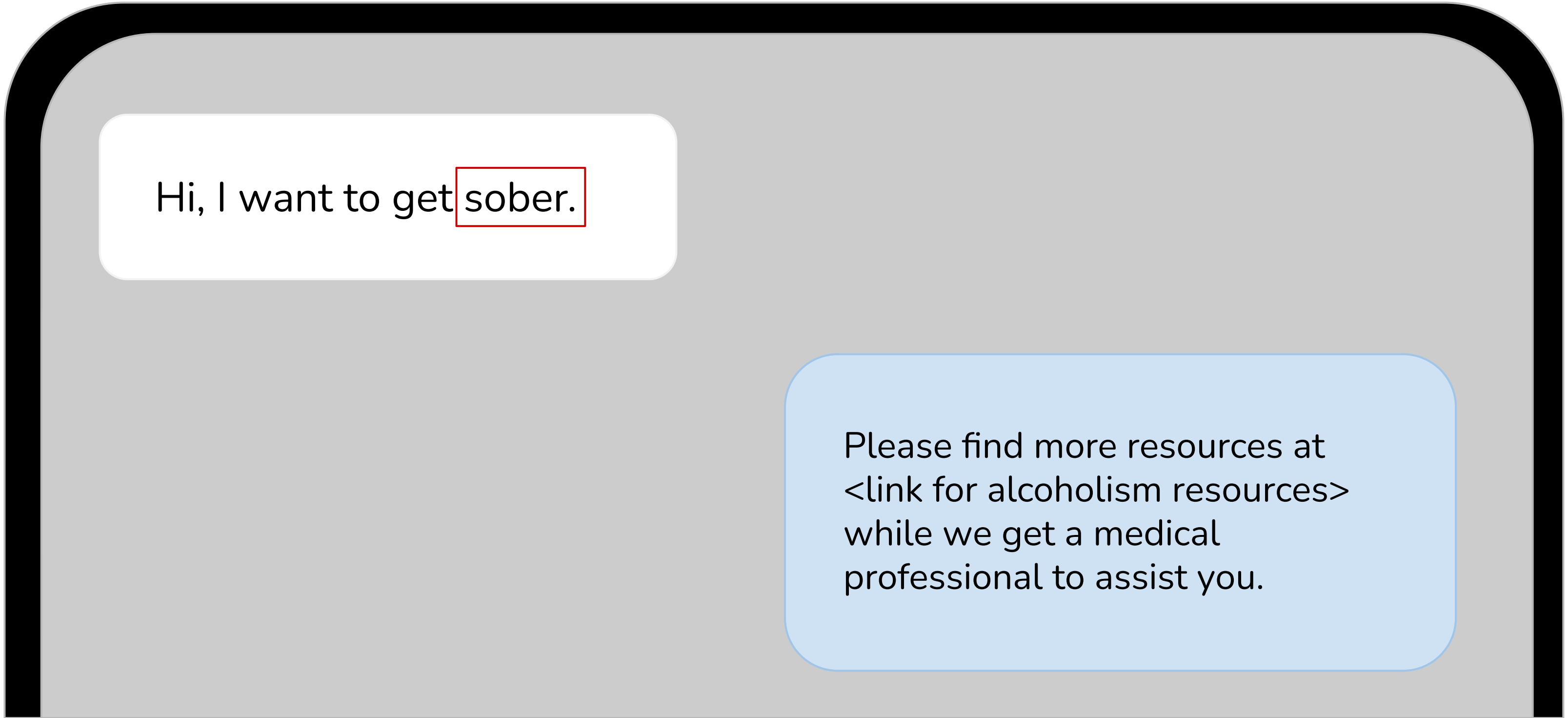
This is a uniquely r/stopdrinking thing that we use as a sign of solidarity. Today I will stay sober, and so will you..it is a statement of support, and a recognition that no one does this alone.

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

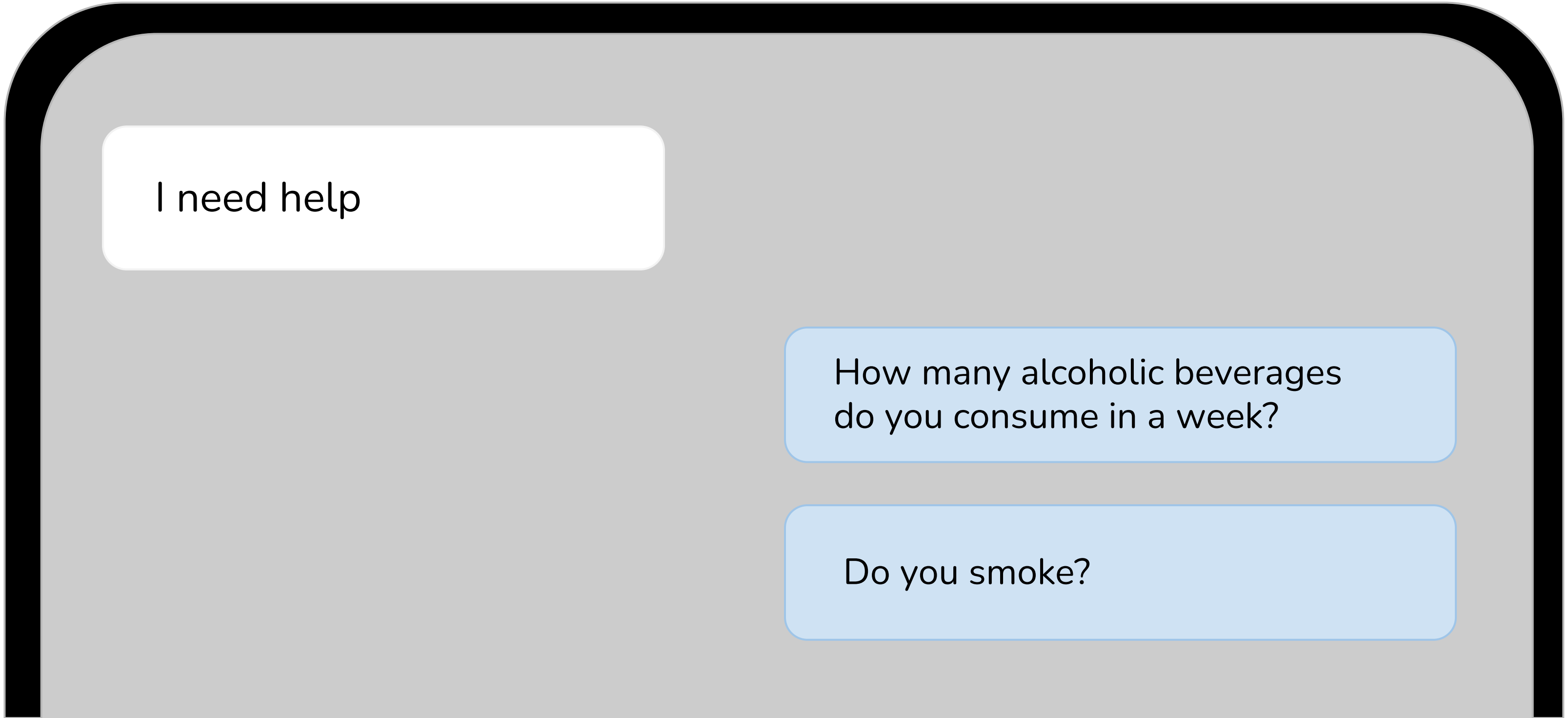
Conclusions & Recommendations



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

Additional Stopwords removed:

['day', 'im', 'like', 'time', 'year', 'feel', 'get', 'one', 'want', 'month',
'really', 'would','go','week','ive','dont','going','thing','even', 'of',
'my' 'meeting', 'people', 'first', 'aa', 'back', 'work', 'much', 'last',
'think', 'got', 'need', 'the', '']