Project 3: Web APIs & NLP

[Internet and Alcohol Addiction Subreddits]

by Timothy Chan

17 Mar 2023

Agenda

- Problem statement
- Approach
- Data scrapped
- Preprocessing
- EDA
- Model and Results
- Error Analysis
- Recommendations

Problem Statement

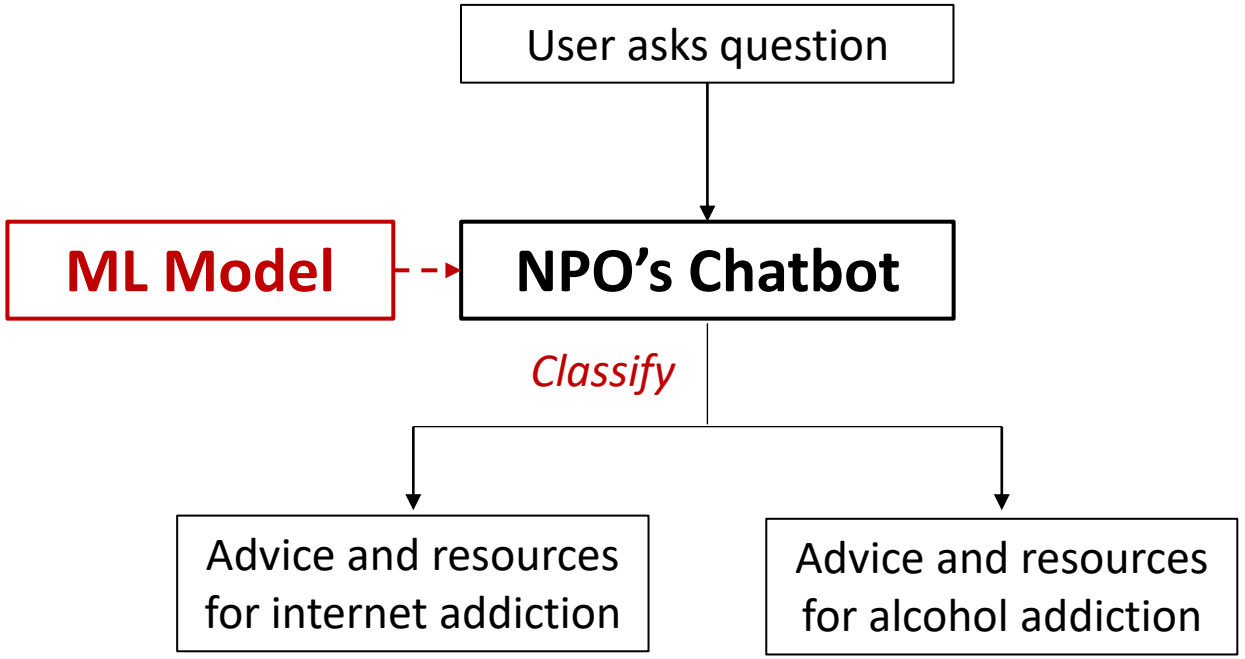
Internet addiction

- Around 8% of global population addicted
- -ve consequences e.g. social isolation and health
- Not yet recognized as disorder by WHO

Alcohol addiction

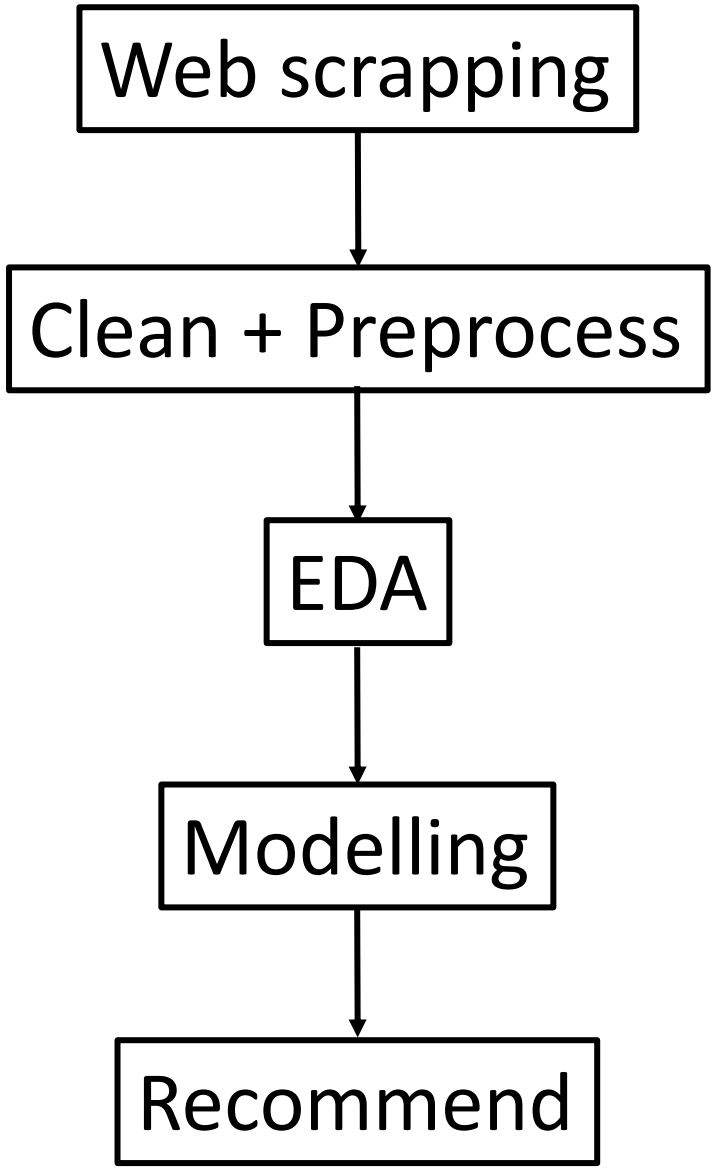
- Limited treatment options
- Impact on individuals, families, and communities
- Over 70% will relapse at some point
- Stigma may prevent them from seeking help

Problem statement

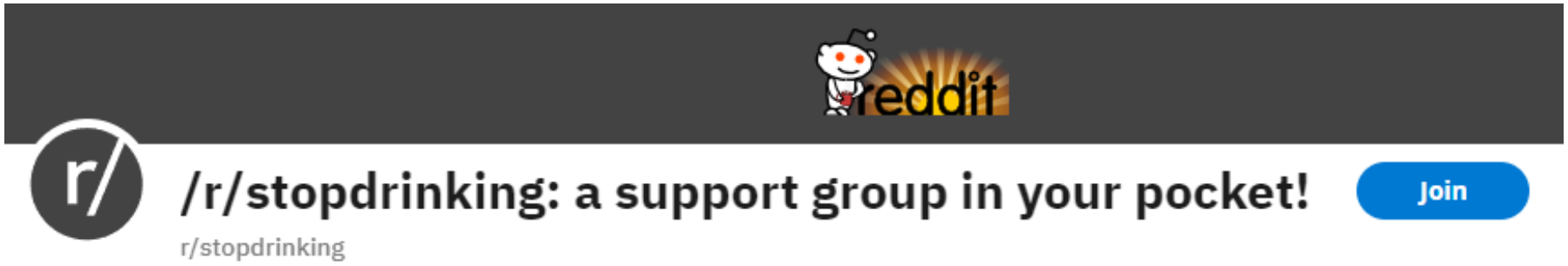


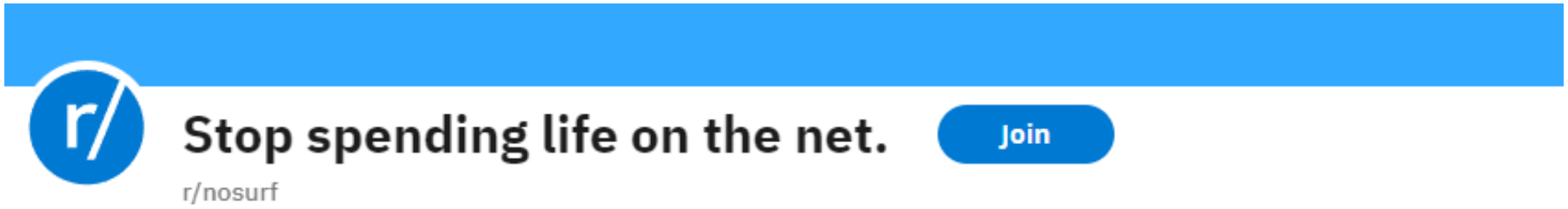
Maximise accuracy, minimal misclassification → Better user experience

Approach



Data scrapped





Preprocessing

- Lowercase
- Remove
  - non-useful websites, links, special characters using Regex
  - obvious phrases (manually)
  - stop words using dictionaries:
    - NLTK
    - Count Vectorizer
- Tokenize by words
- Lemmatize

Automate as much as possible

Exploratory Data Analysis

Sentiment Coefficient (NLTK)

r/stopdrinking:

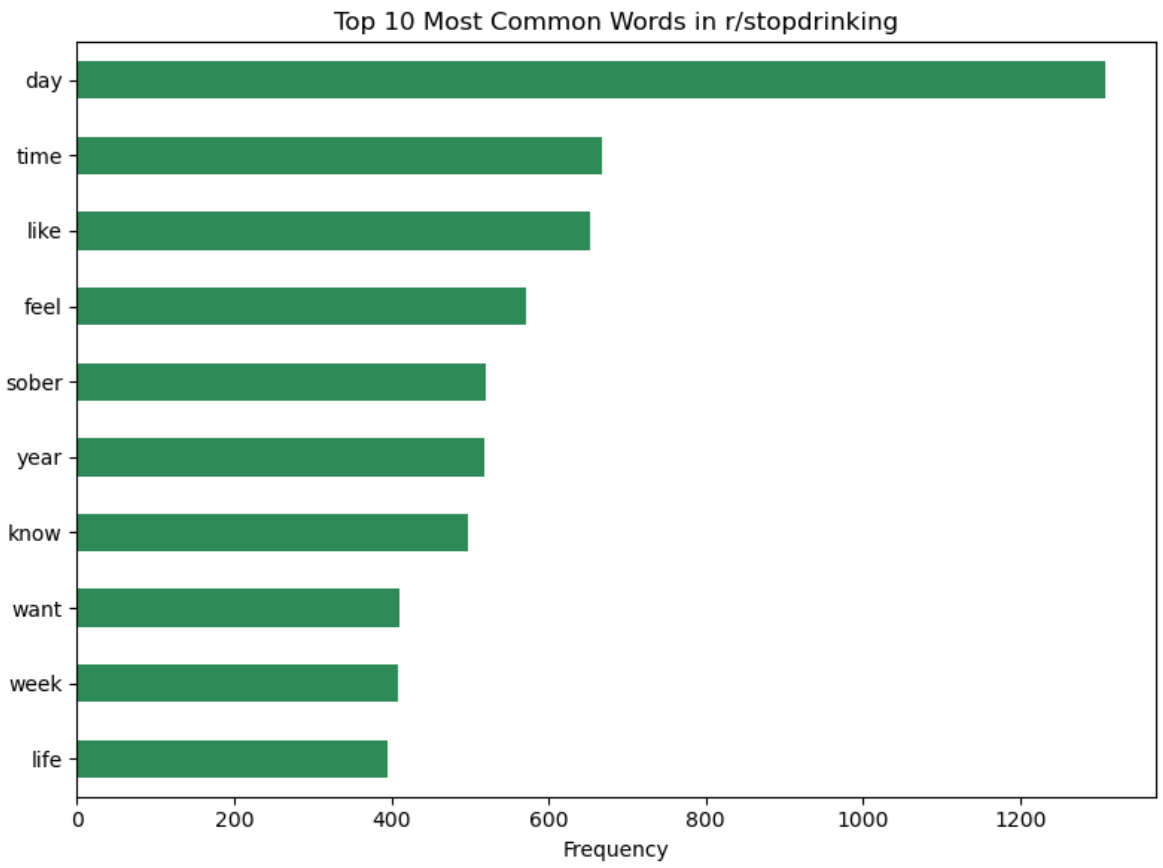
0.257

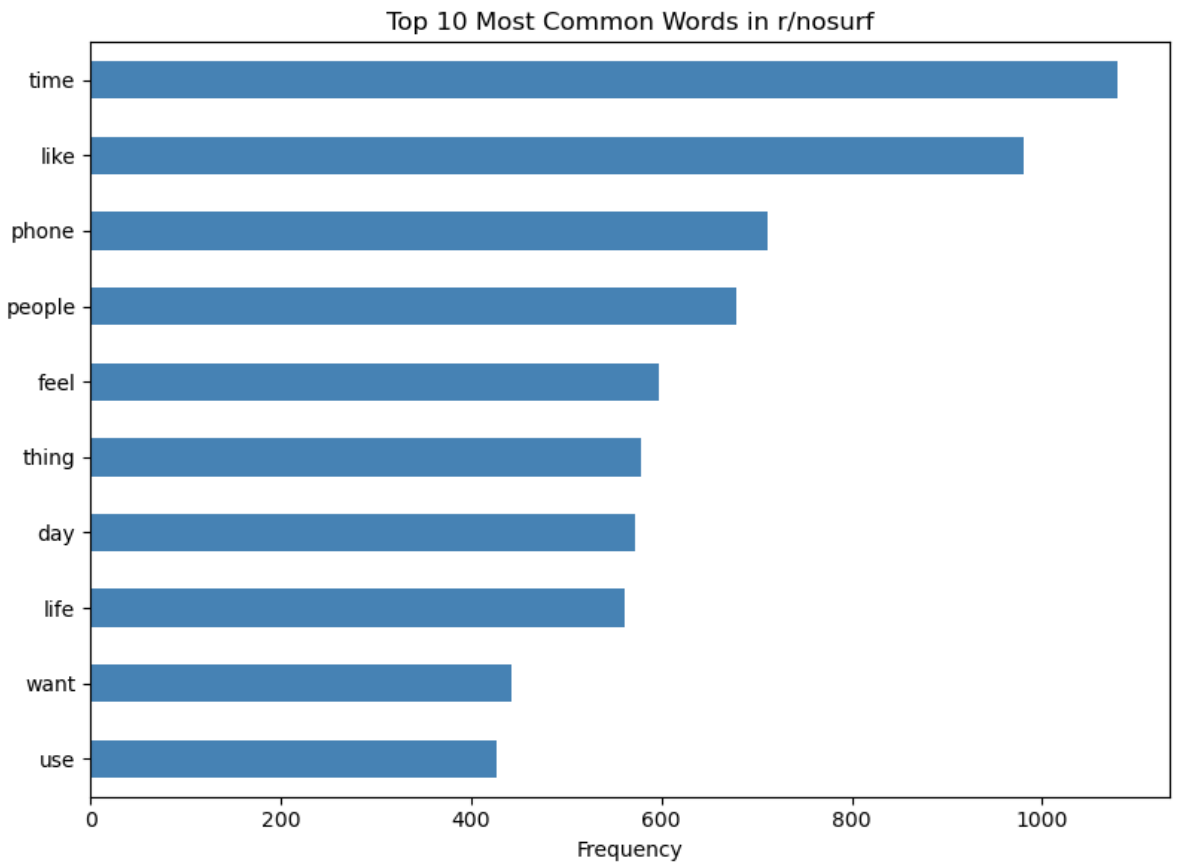
r/stopdrinking:

0.248

Both moderately positive
Will not classify based on sentiments

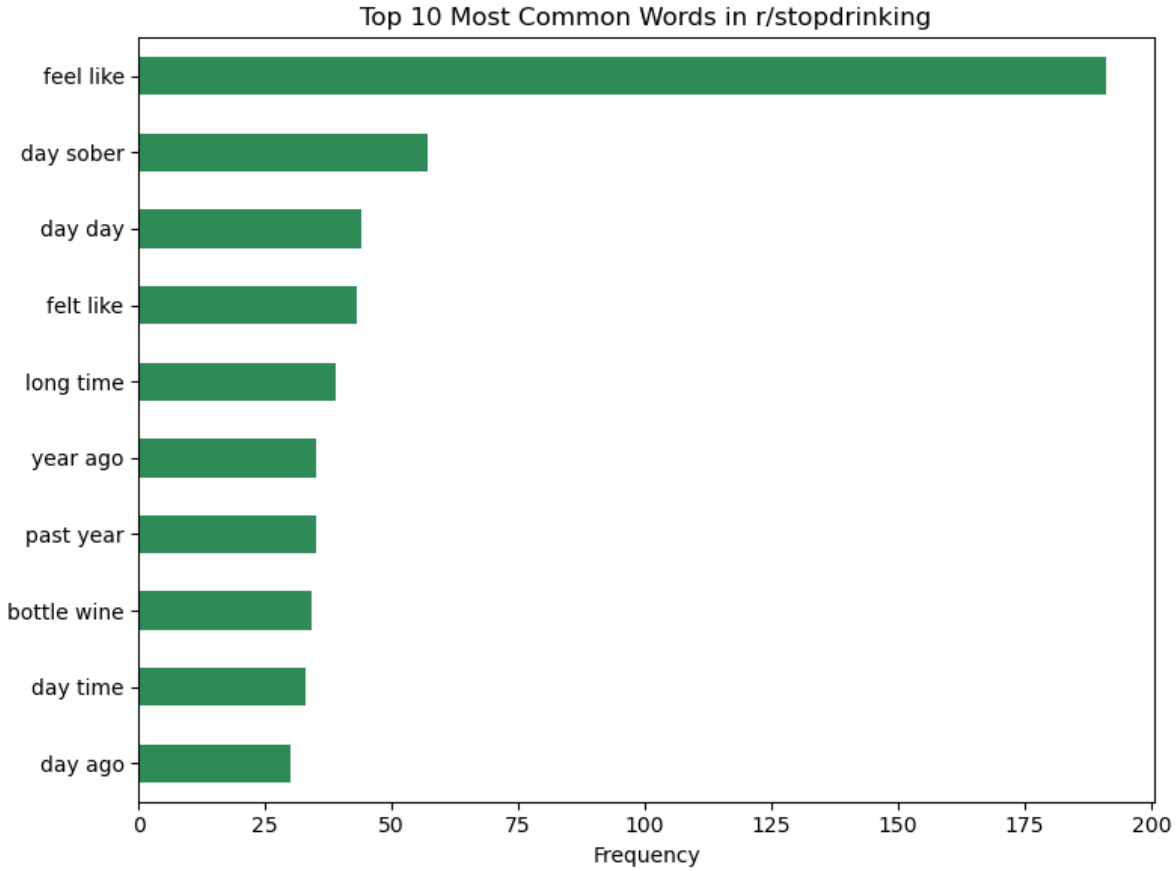
Frequent one word phrase

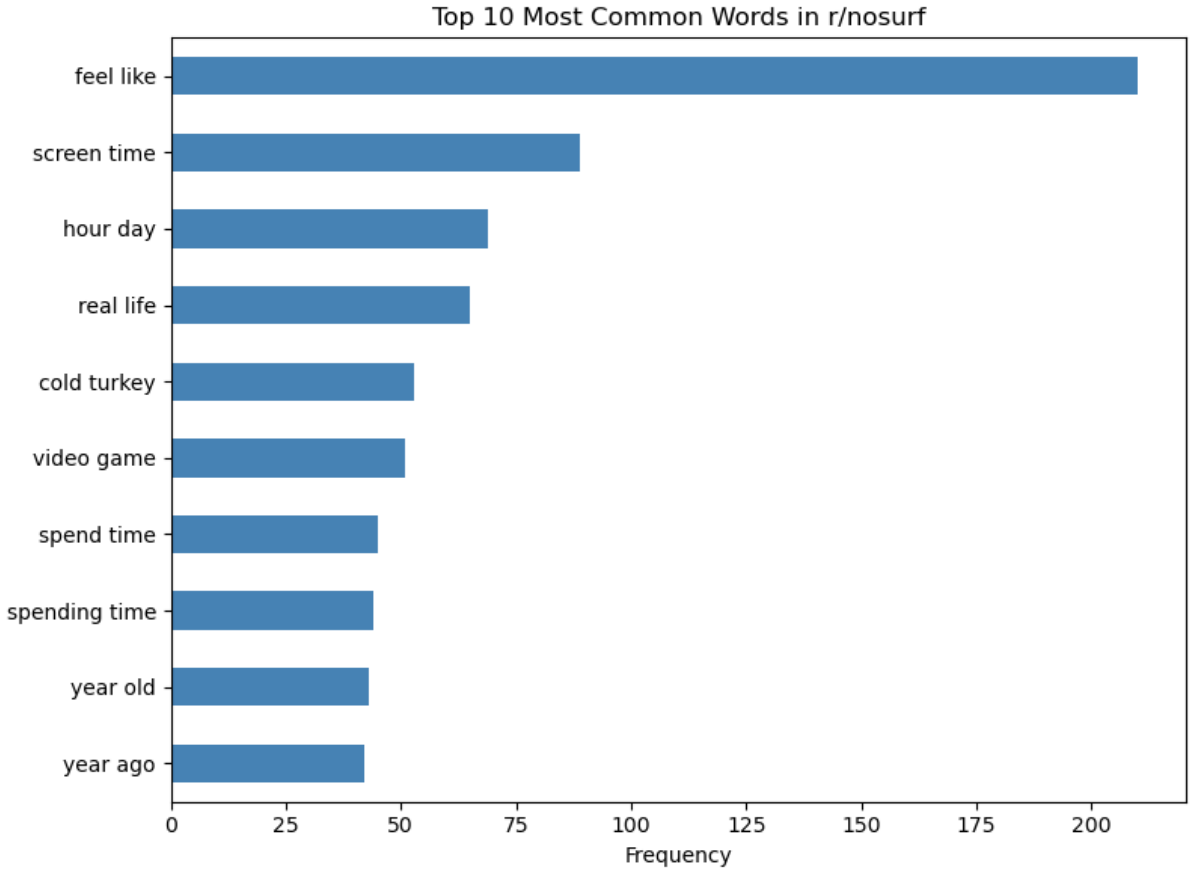




Not too many unnecessary or meaningless words which has not been preprocessed earlier

Frequent two words phrase





Not too many unnecessary or meaningless words which has not been preprocessed earlier

Models and Results

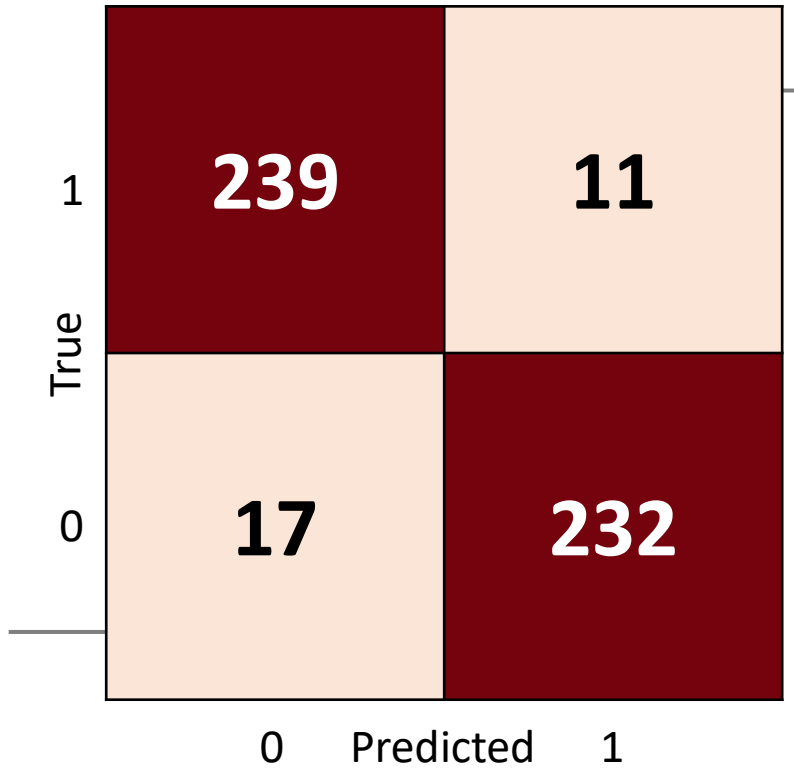
	Count Vectorizer	TF-IDF Vectorizer
Multinomial Naive Bayes	**0.944** *(-0.019)*	0.936 *(-0.040)*
Logistic Regression	0.934 *(-0.047)*	0.932 *(-0.055)*
Random Forest	0.924 *(-0.035)*	0.940 *(-0.025)*
Gradient Boosting	0.920 *(-0.057)*	0.918 *(-0.079)*

* test accuracy score and overfitting (test minus train)

Selected model:

Count Vectorizer x Multinomial Naive Bayes

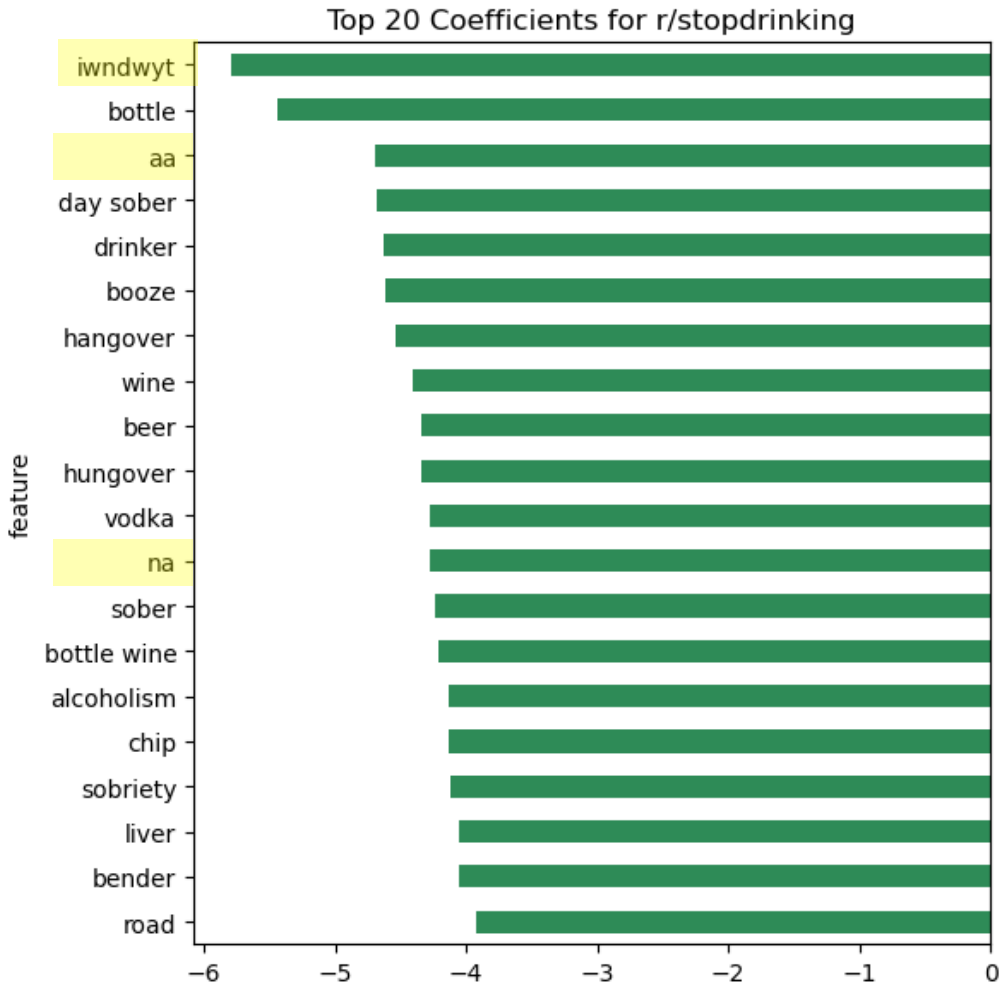
Confusion matrix

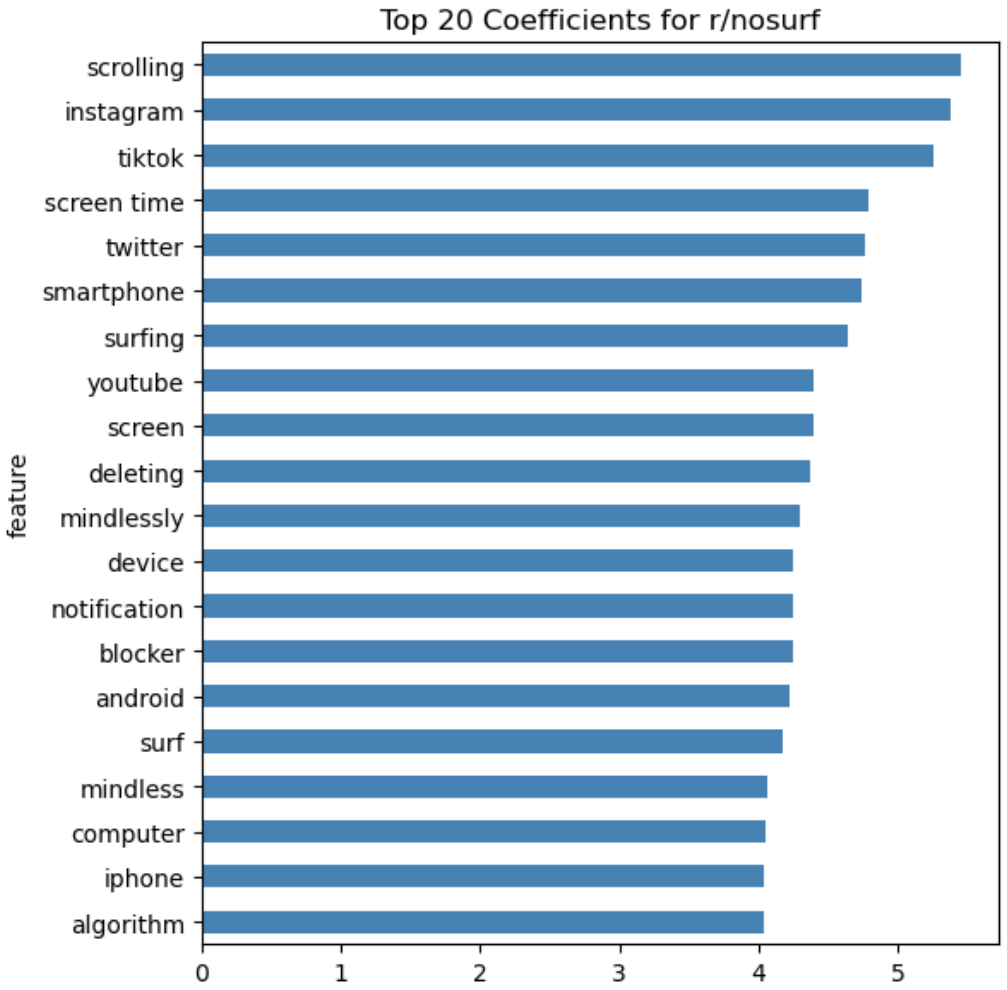


False positive: *r/stopdrinking* wrongly classified as *r/nosurf*

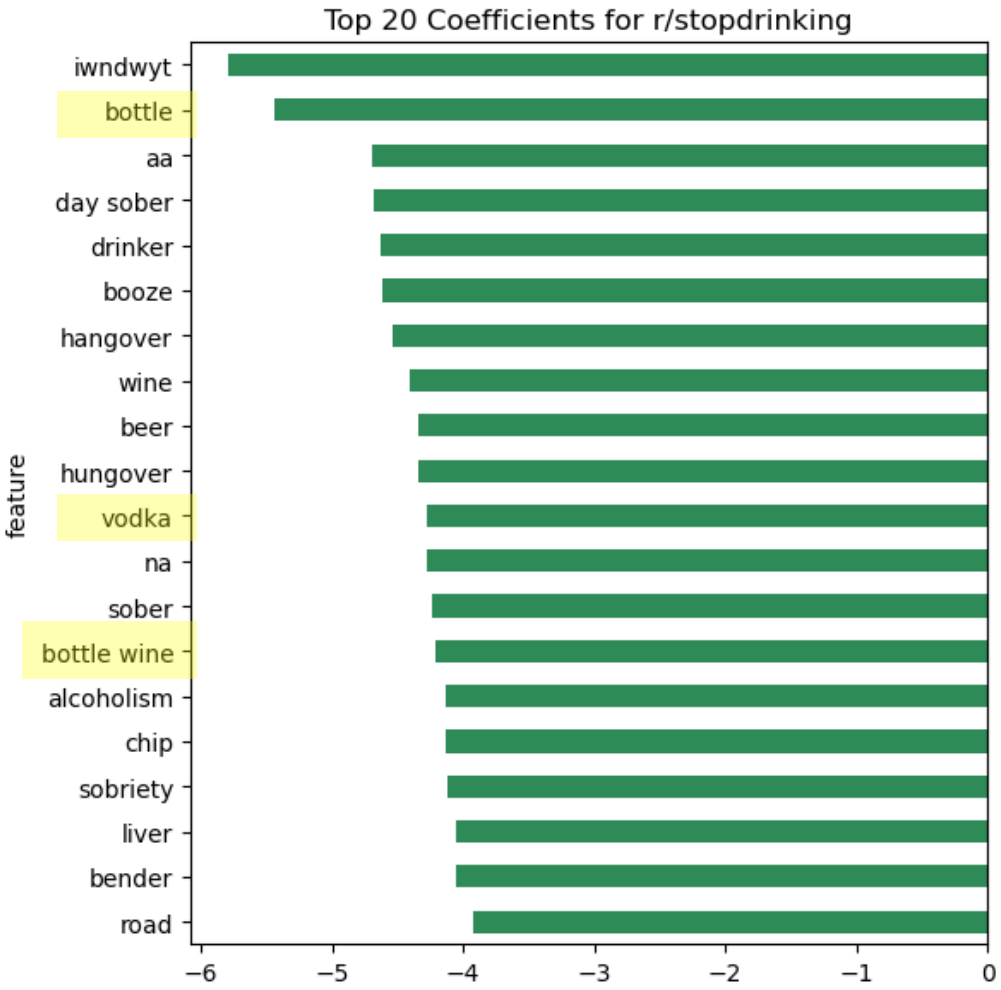
False Negative: *r/nosurf* wrongly classified as *r/stopdrinking*

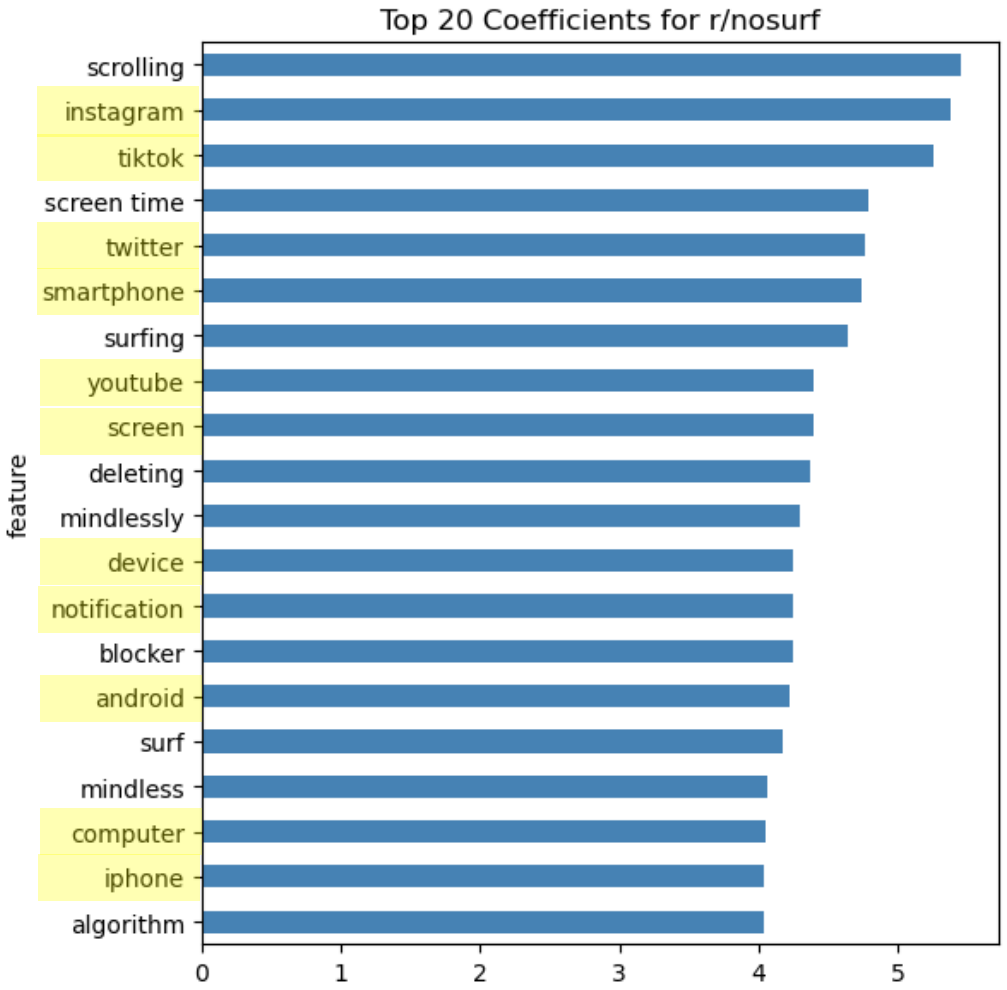
Top 20 coefficients: no abnormal phrases



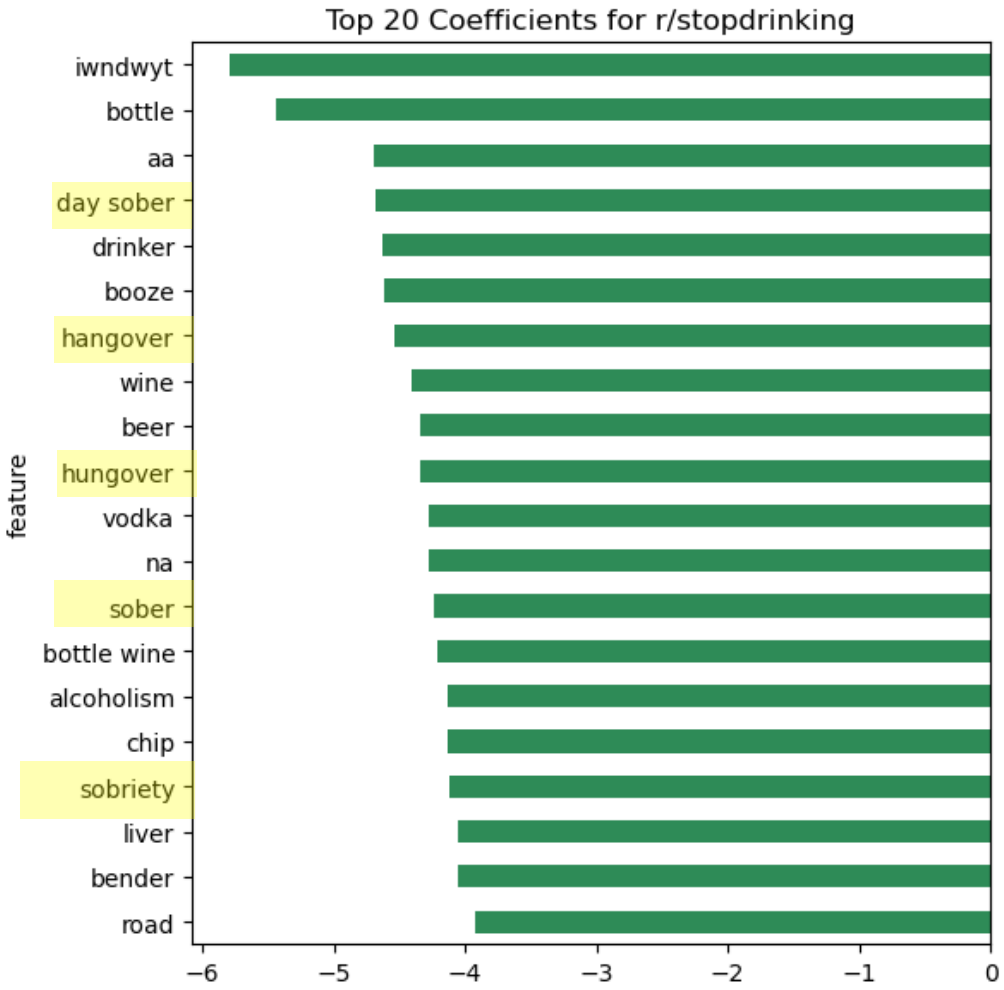


Top 20 coefficients: addiction medium

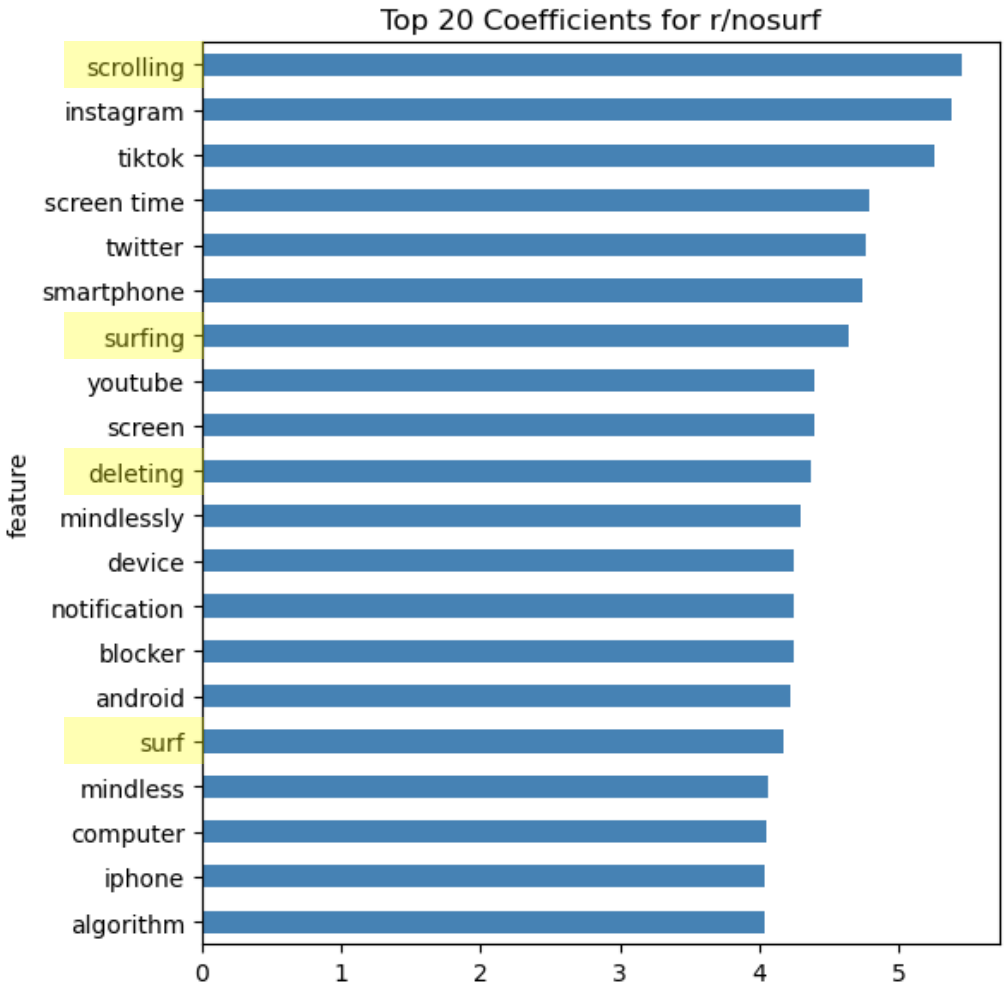




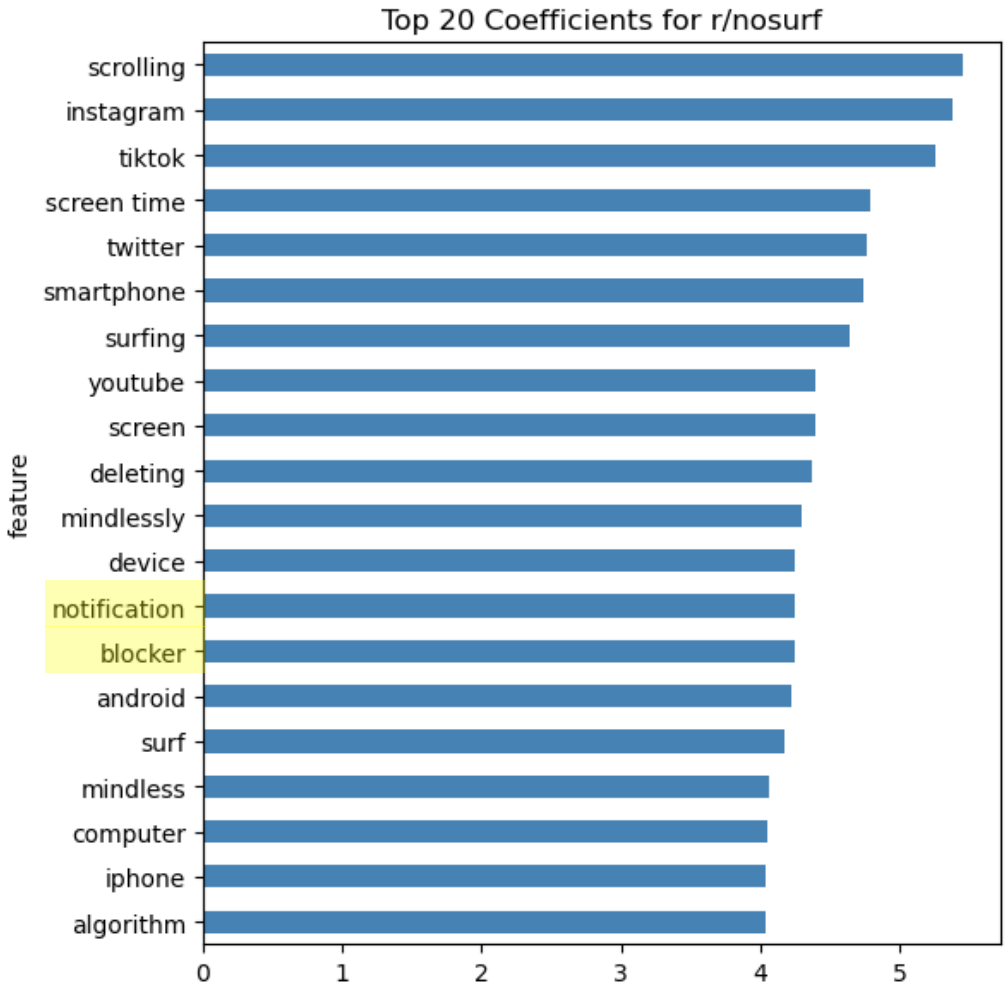
Top 20 coefficients: state of being drunk



Top 20 coefficients: actions



Top 20 coefficients: means to change habits



Error Analysis

False Negative: r/nosurf wrongly classified as r/stopdrinking

r/nosurf · Posted by u/endthenet 1 month ago

11

Scorched earth - Day 3

The silence is deafening. Has it always been this quiet? This is the first time that I feel so bored that I actually want to work. Anything is better than just sitting here doing nothing.

5 Comments Share Save Hide Report 100% Upvoted

Posted by u/[deleted] 18 days ago

2

Doomscrolling and being a minority

Sorry, this post was deleted by the person who originally posted it.

It doesn't appear in any feeds, and anyone with a direct link to it will see a message like this one.

False Positive: r/stopdrinking wrongly classified as r/nosurf

Posted by u/guysweepingstreet 15 days 12 days ago

3

Anyone else using the NoMo sobriety day counting app?

I installed NoMo last night, it is free and can actually be used for all kinds of addiction cessation. There are features besides the counter and I wondered if anyone has used them.

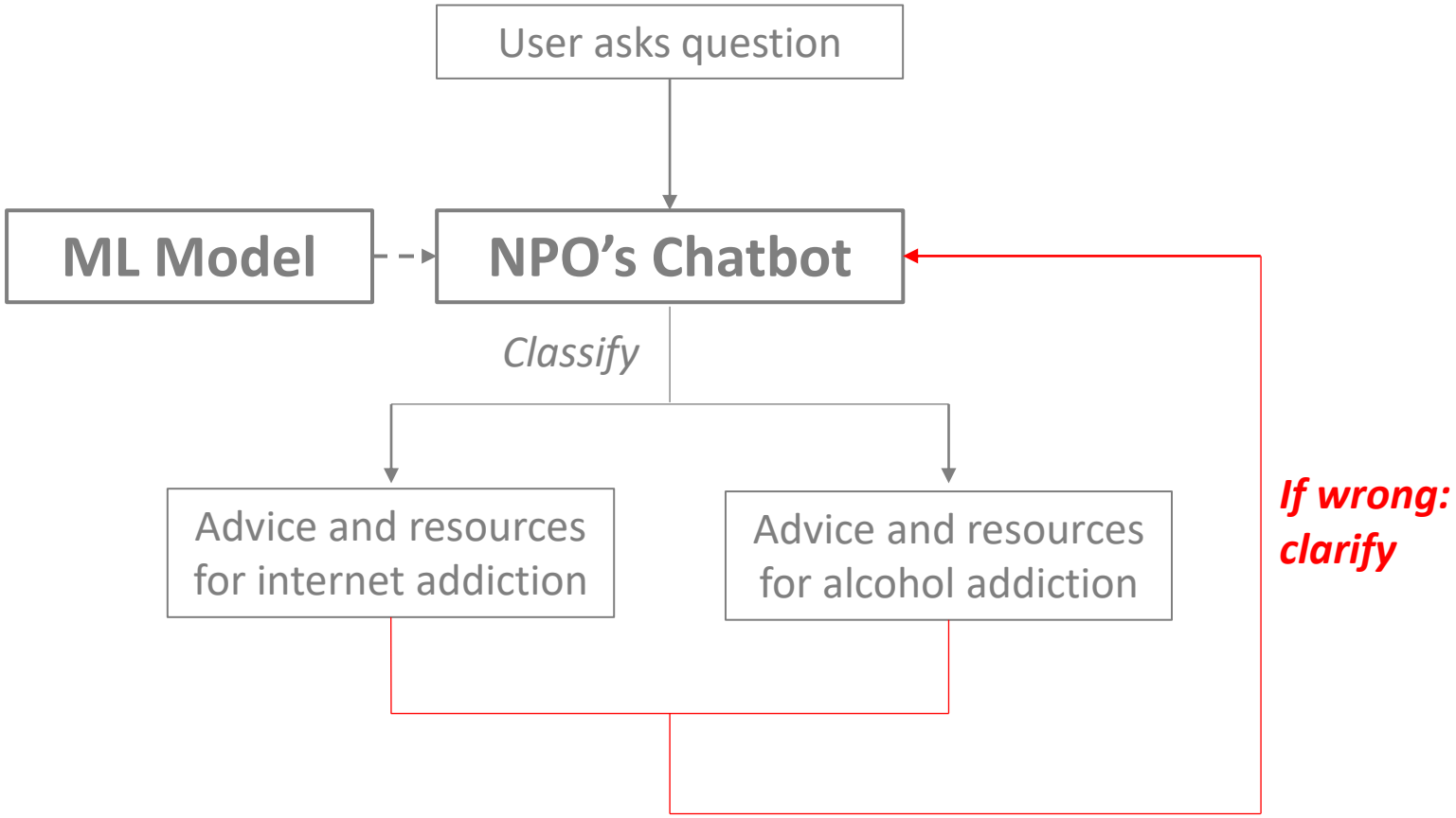
Posted by u/hey-now_easy-now 13 days ago

8

Alcohol is a liar

Check out Henry Rollins, music video "Liar" to remind you not to believe the nice things alcohol is trying to tell you. (No link because video links are not allowed here)

When there are errors



Recommendations

- Count Vectorizer with Multinomial NB: high accuracy of 0.944
- False results mostly input issues and not model issue
- Higher accuracy means better user experience
- However, not critical (and practical) for chatbot to be fully accurate
- Rectification through chatbot user clarification in this case
- Automated preprocessing as much as possible

Possible improvements

- More data to improve accuracy further
- Multiple sources besides Reddit for better reflection of words to be trained
- Finetuning of hyperparameters, especially for ensemble models
  - to improve performance and/or reduce overfitting
  - however, this requires more time or resources
- Extend to other types of addictions or more targeted advice based on needs
  - more complex model

Questions