# Whiskey vs Rum

Woo Junhong

# Table of contents

# 01 Background

# Sing Song Cellar

- International alcohol dealer
- Good quality of alcohol
- First company to air drop alcoholic beverages
- Affordable alcohol

# Problem statement

Creation of a machine learning model to maximize the efficiency of their marketing spend based on their target audience.  Whiskey or Rum?

# Goals

Aim To build a classifier that identifies keywords and to classify them to either whiskey or rum based on accuracy

# WEB-SCRAPING

- PushShift API
- At least 10,000 rows

r/rum

r/whiskey

# Data Cleaning

- Combined title and texts features
- Combined into one dataframe
- Duplicates removed

# Preprocessing text

01 Remove links

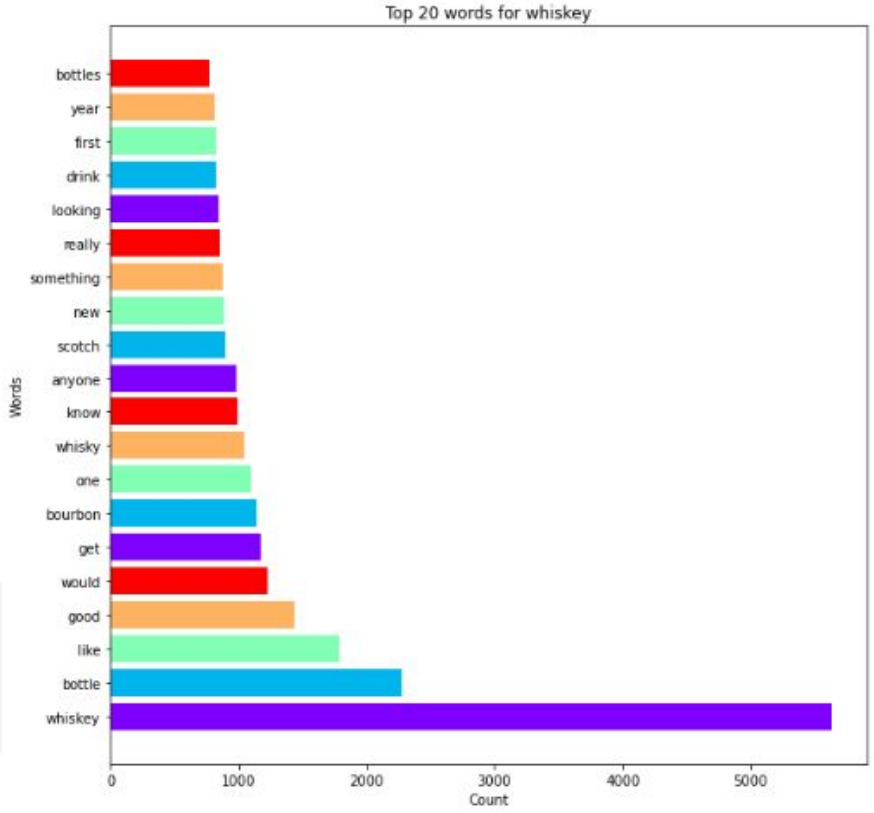
02 Remove special characters, emoji

03 Remove emojis

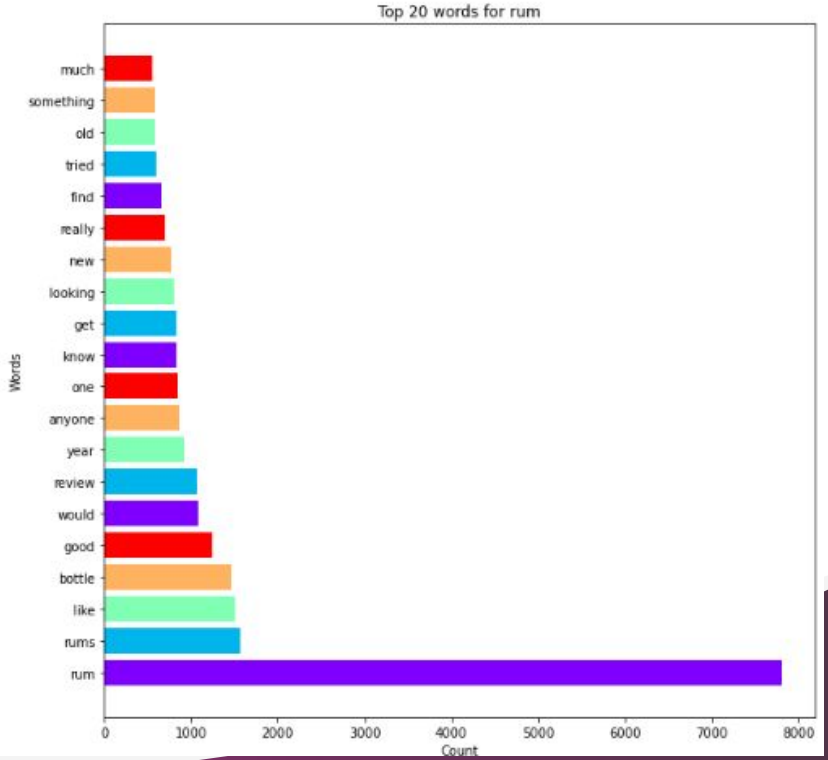
04 Tokenizer and lemmatizer

# Top Words

## r/whiskey

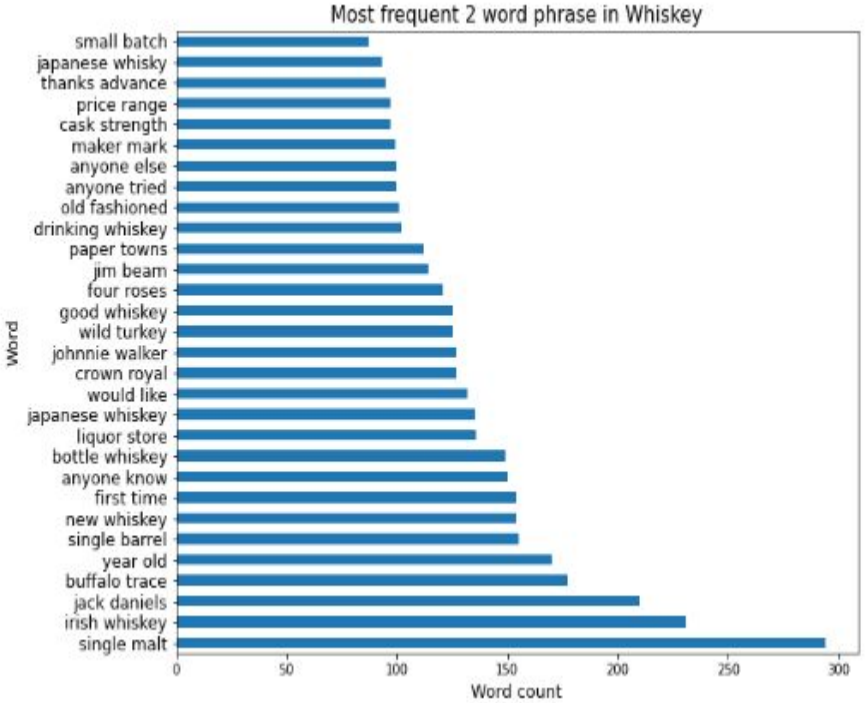



## r/rum

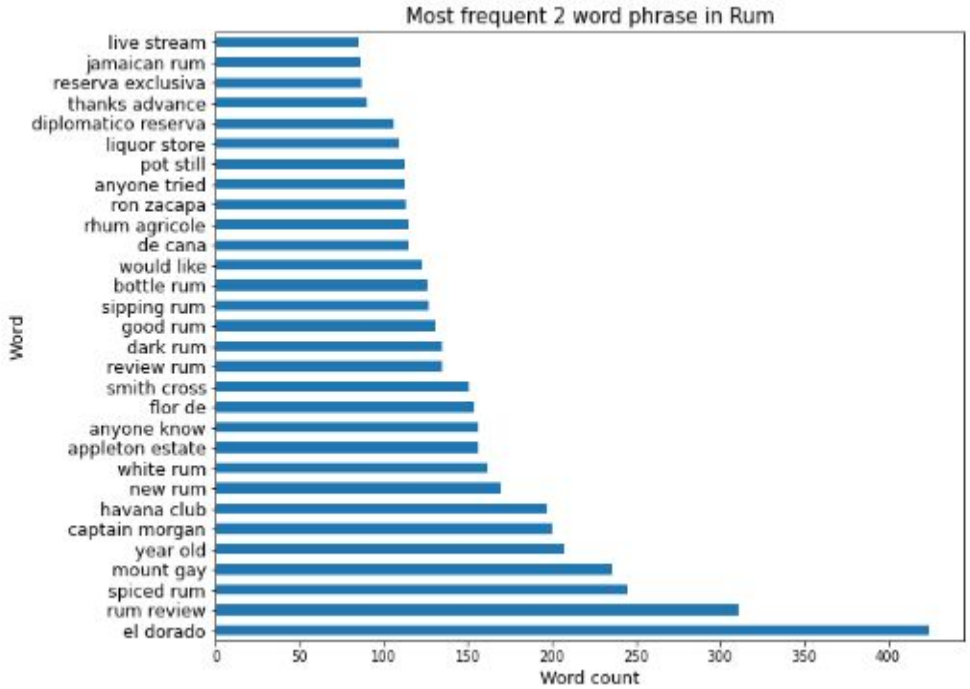


# Top 2 word phase

## r/whiskey

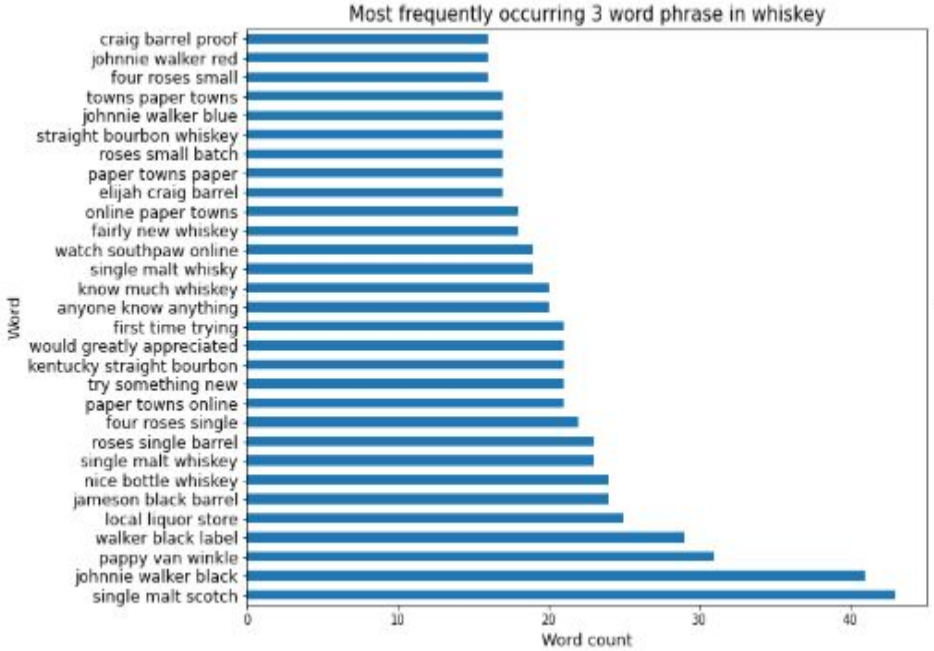



## r/rum

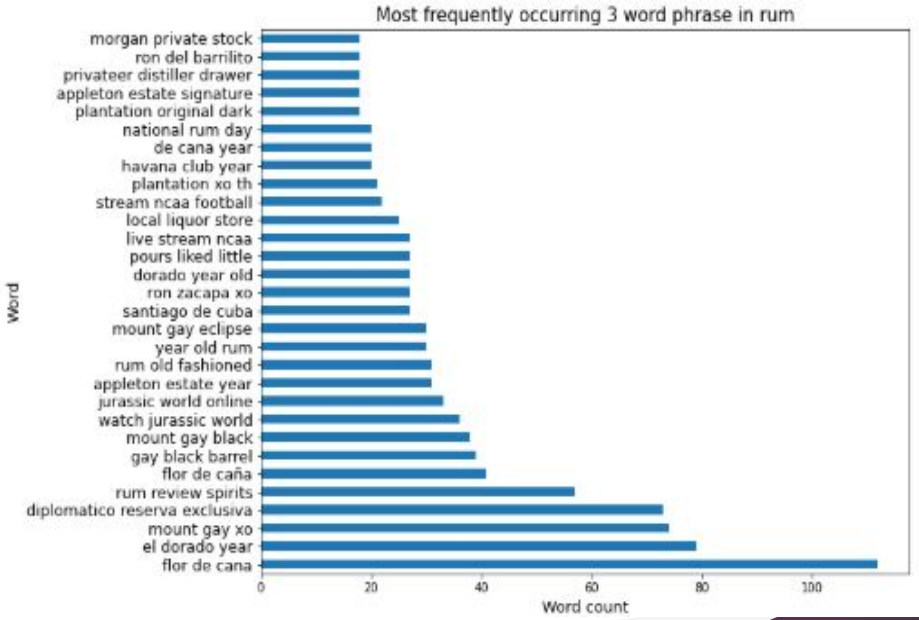


# Top 3 word phase

## r/whiskey

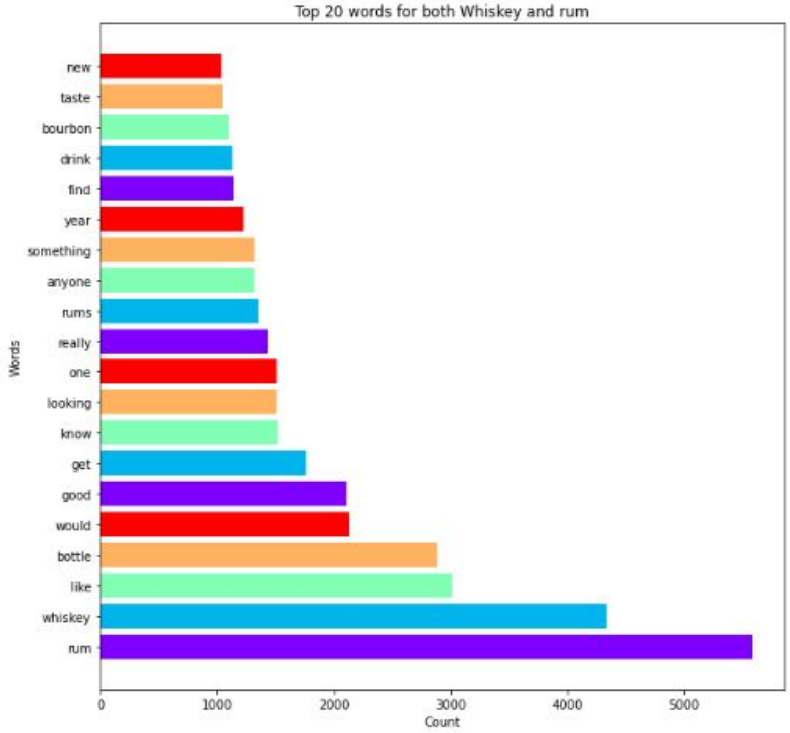


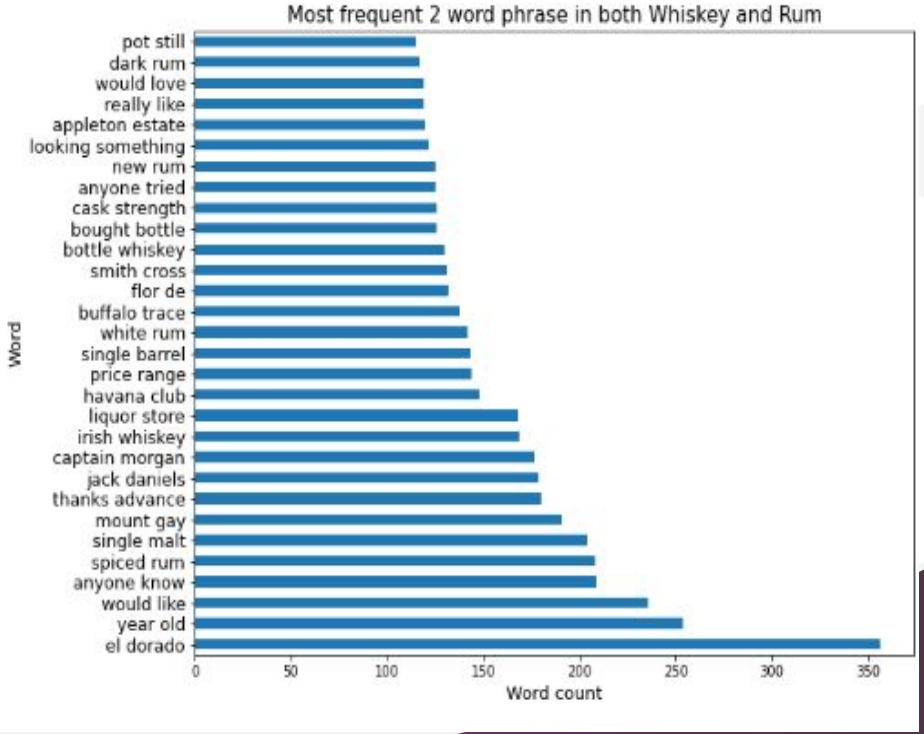

## r/rum



# Rum and whiskey top words

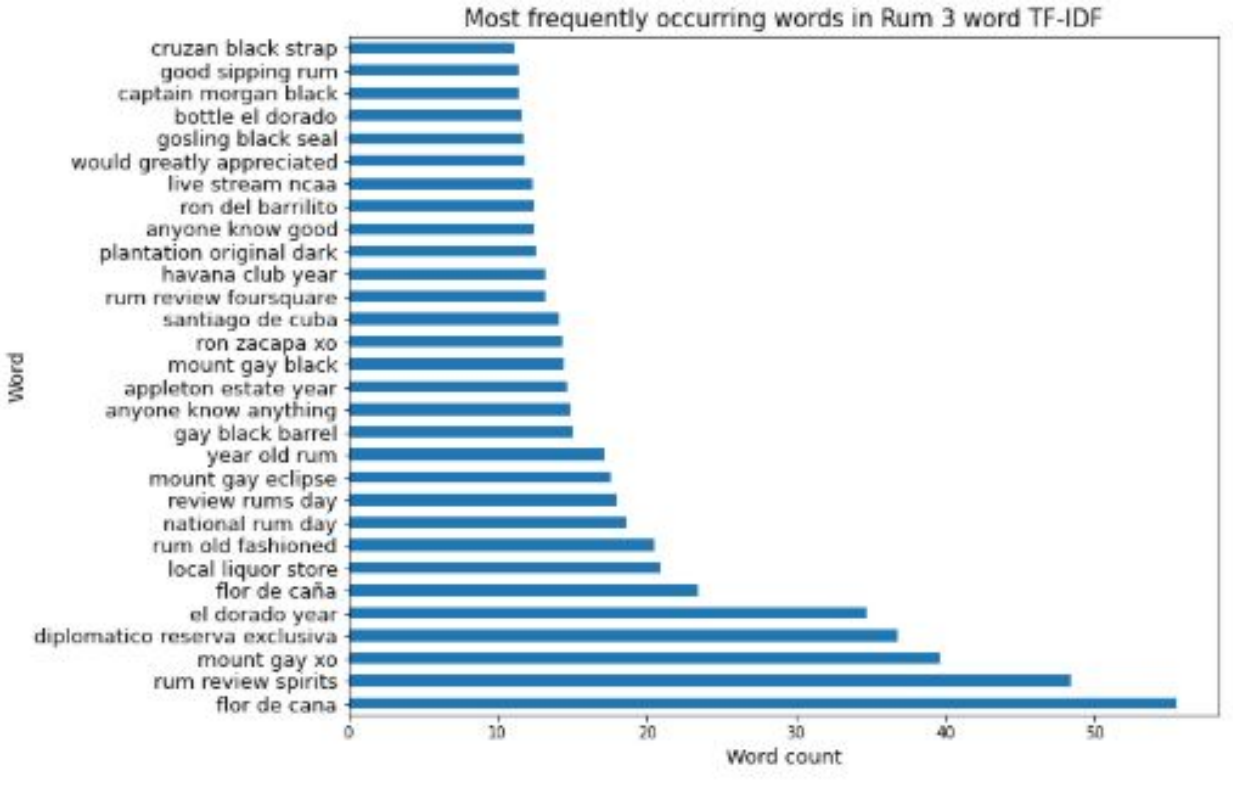
## For Rum and whiskey



# Rum and whiskey top words

## For Rum and whiskey



# Modeling and Evaluation

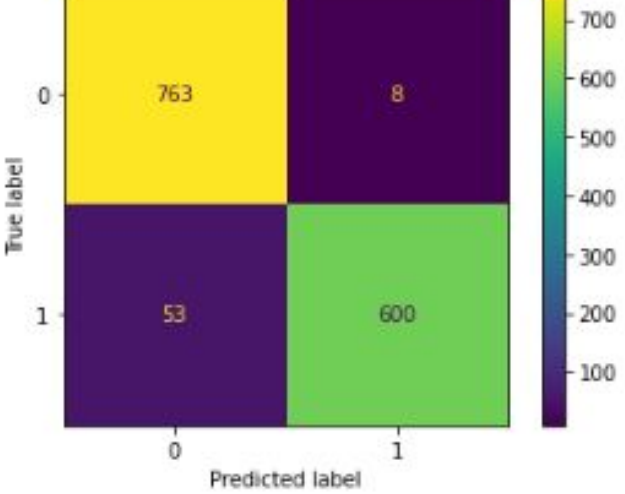
## Baseline accuracy : 0.54

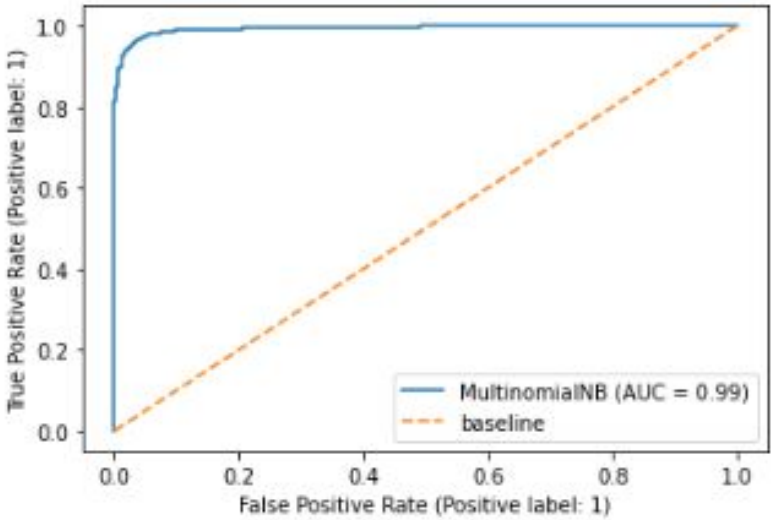
| Model | Train Score | Test Score |
|---|---|---|
| Bernoulli Naive Bayes (TfidfVectorizer) | 0.54 | 0.54 |
| Bernoulli Naive Bayes (CountVectorizer) | 0.88 | 0.86 |
| Multinomial Bayes (CountVectorizer) | 0.88 | 0.86 |
| Multinomial Bayes (TF-IDFVectorizer) | 0.98 | 0.95 |
| Gaussian Naive Bayes (CountVectorizer) | 0.88 | 0.86 |
| Gaussian Naive Bayes (TfidfVectorizer) | 0.98 | 0.80 |
| Logistic Regression (TF-IDFVectorizer) | 0.99 | 0.96 |
| KNeighborsClassifier (TF-IDFVectorizer) | 0.94 | 0.89 |
| Hypertuned KNeighborsClassifier (TF-IDFVectorizer) | 0.89 | 0.88 |

# Model Fitting

Multinomial NB model - TFIDF vectortizer

| | | |
|---|---|---|
| Multinomial Bayes (TF-IDFVectorizer) | 0.98 | 0.95 |




# Model Fitting

- AUC of 0.99
- Better at distinguishing positive and negatives

# Recommendations

## Explore other models

Random Forests Classifier, Ensemble Techniques, etc

## Add more stop words

Use more useful words

## Hypertune models

Both vectorizers and model to compare

## Limited to 2 subreddits

Include more subreddits into model

## More data

Other social media platforms