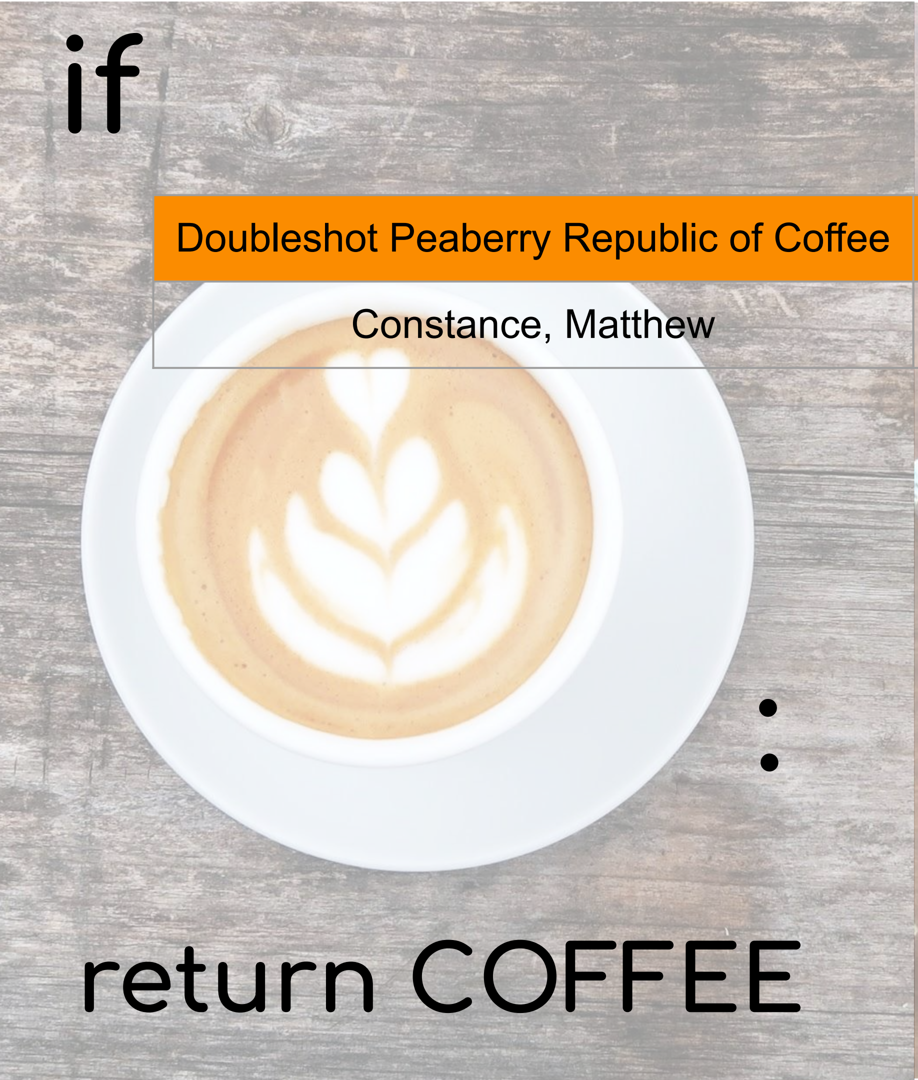
DSI-41 Project 3



if :

else :

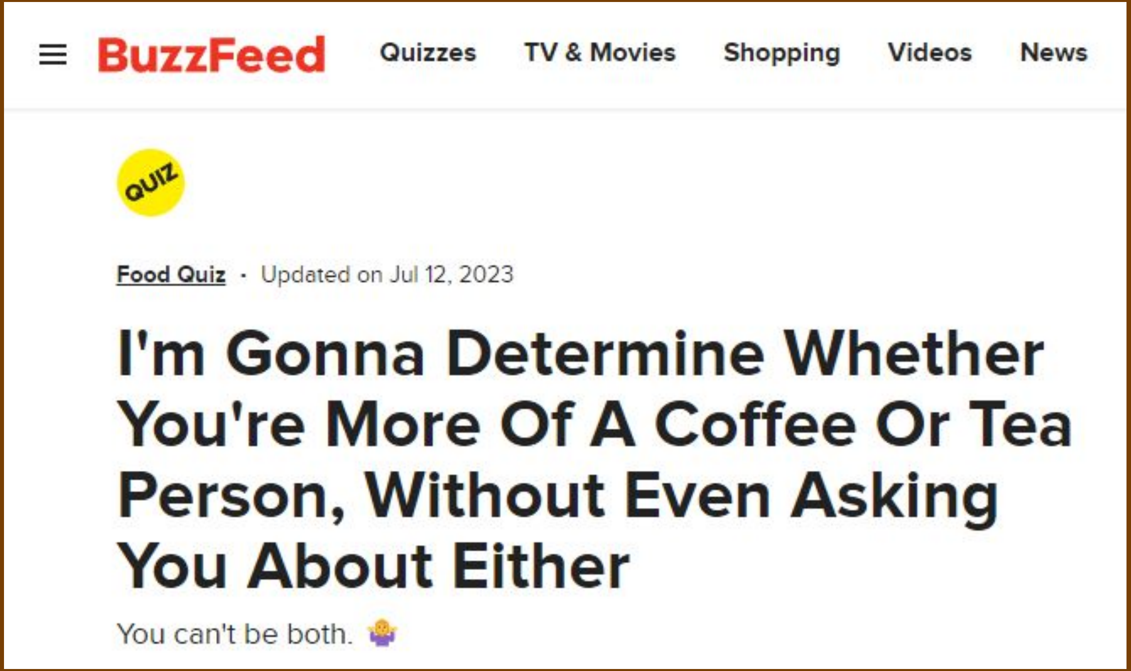
Doubleshot Peaberry Republic of Coffee	Decaf Pu-Erh Republic of Cha
Constance, Matthew	Joel, Wenzhe

return COFFEE

return TEA

Choose a side!

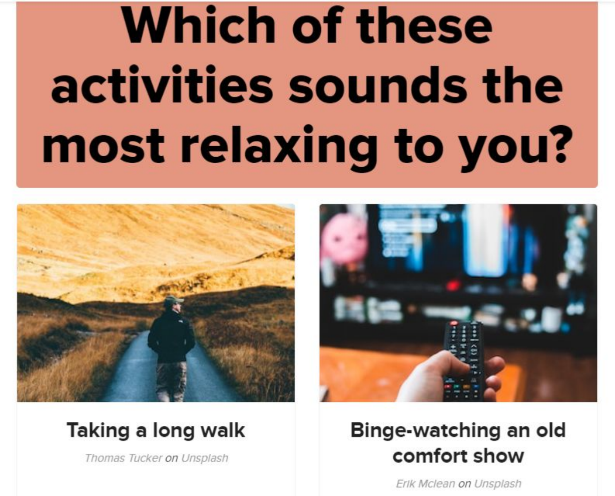
https://www.buzzfeed.com/sofia111/coffee-or-tea-person-personality-quiz



Food Quiz · Updated on Jul 12, 2023

I'm Gonna Determine Whether You're More Of A Coffee Or Tea Person, Without Even Asking You About Either

You can't be both. 🤷‍♀️



Which of these activities sounds the most relaxing to you?

Taking a long walk

Thomas Tucker on Unsplash

Binge-watching an old comfort show

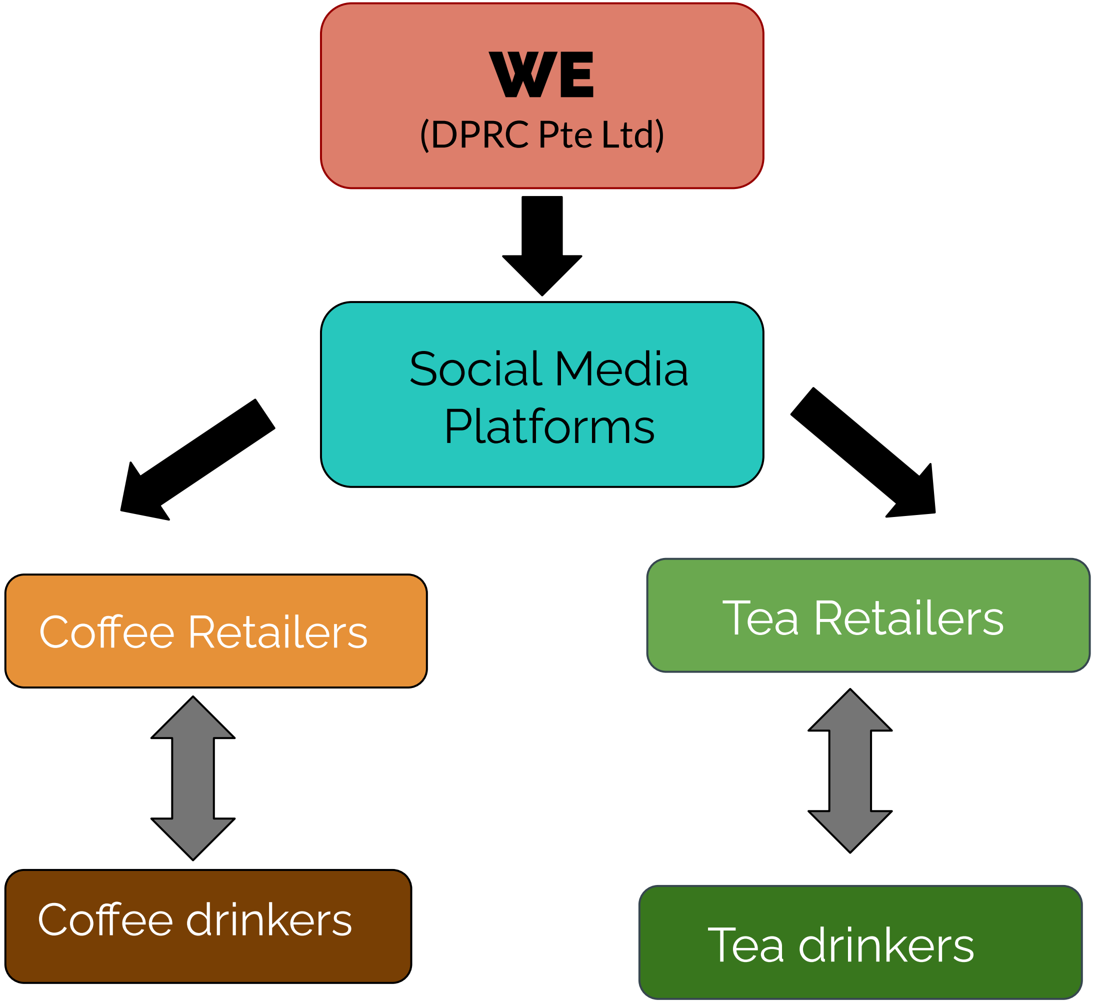
Erik Mclean on Unsplash

Problem Statement

What are we trying to solve?

To help social media platforms to perfect their targeted ads based on user comments for their F&B clients

Our Value Proposition



WE

(DPRC Pte Ltd)

Social Media Platforms

Coffee Retailers

Tea Retailers

Coffee drinkers

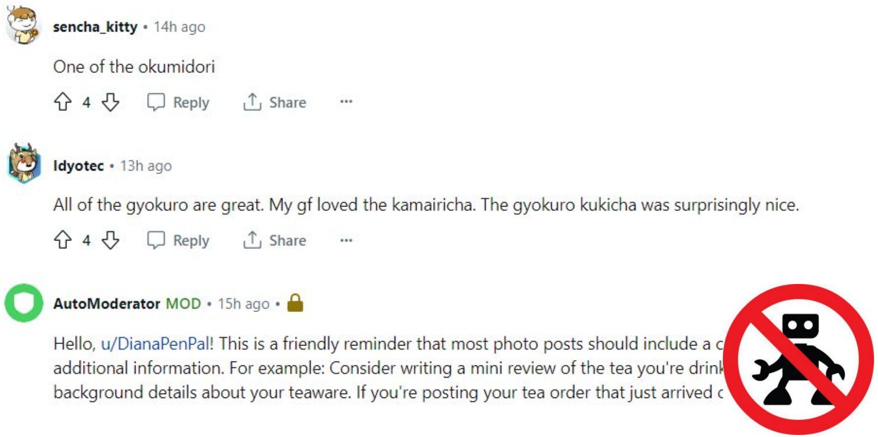
Tea drinkers

Increase the ability of social media platforms to target coffee/tea drinkers

1. Allow them to **increase attractiveness**
2. **Revenue**

Data Scraping

What did we do?



sencha_kitty • 14h ago

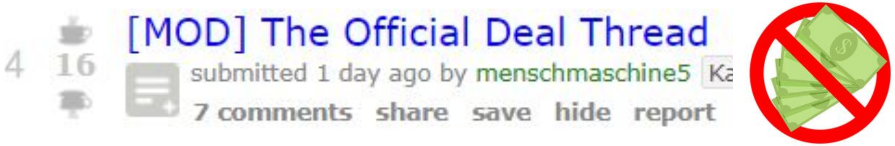
One of the okumidori

Idyotec • 13h ago

All of the gyokuro are great. My gf loved the kamairicha. The gyokuro kukicha was surprisingly nice.

AutoModerator MOD • 15h ago •

Hello, u/DianaPenPal! This is a friendly reminder that most photo posts should include a c
additional information. For example: Consider writing a mini review of the tea you're drink
background details about your teaware. If you're posting your tea order that just arrived c



[MOD] The Official Deal Thread

submitted 1 day ago by menschmaschine5 Ka

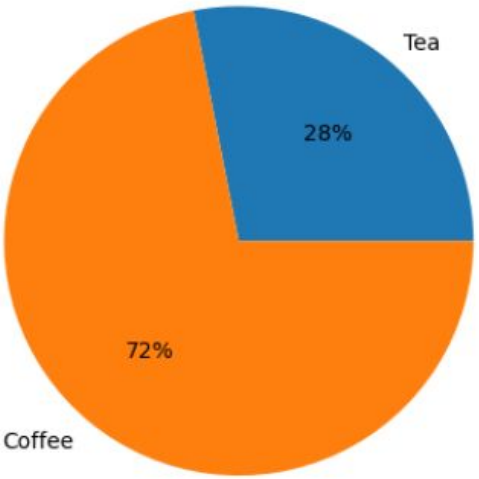
7 comments share save hide report

- ➔ ~**1000 "hottest" threads** from r/coffee and r/tea
- ➔ Each comment in reply to the main thread is stored
- ➔ **Filter marketing threads** from both subreddits
- ➔ **Further data cleaning** to remove links, and posts by bots

Data Sampling

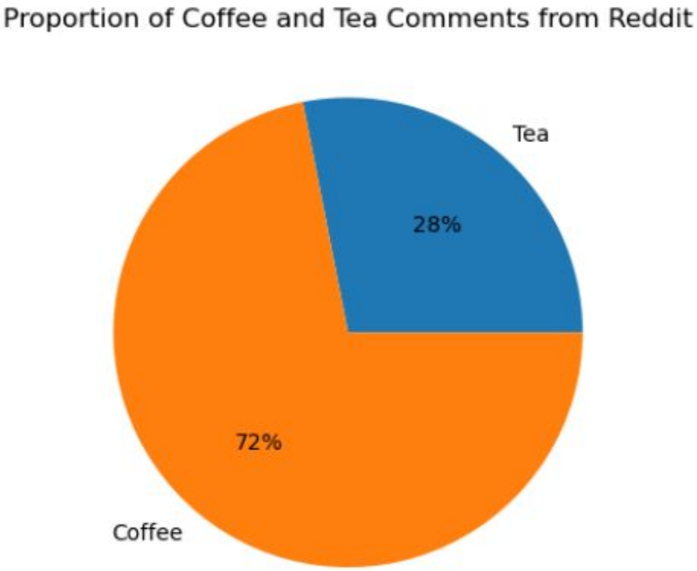
Before Sampling

Proportion of Coffee and Tea Comments from Reddit

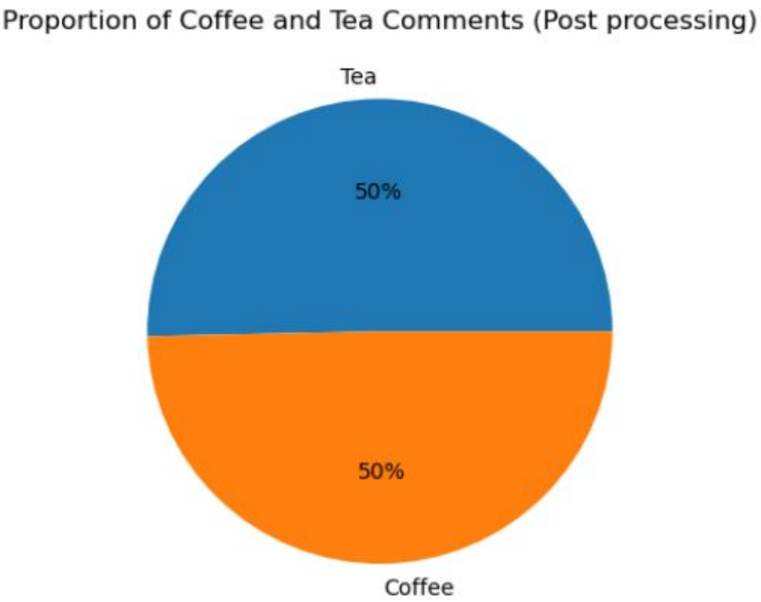


- ➔ Scores tend to indicate the level of agreement of users in the subreddits; hence, we want to **capture the different sentiments**
- ➔ We choose to retain unique comments from each thread with **highest and lowest scores**

Before Sampling



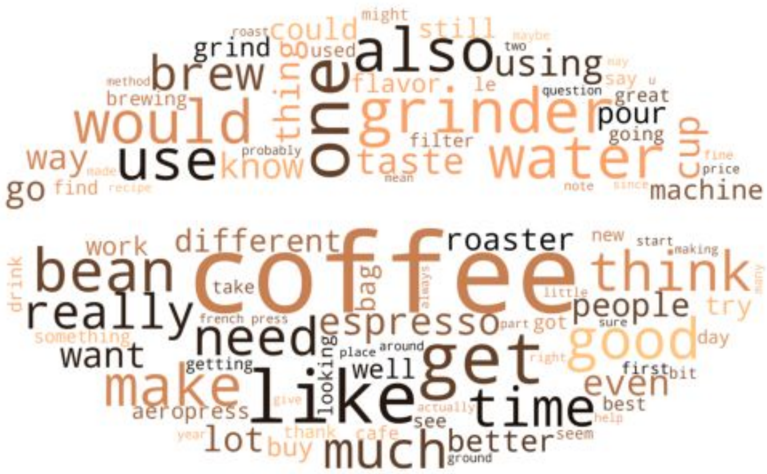
After Sampling

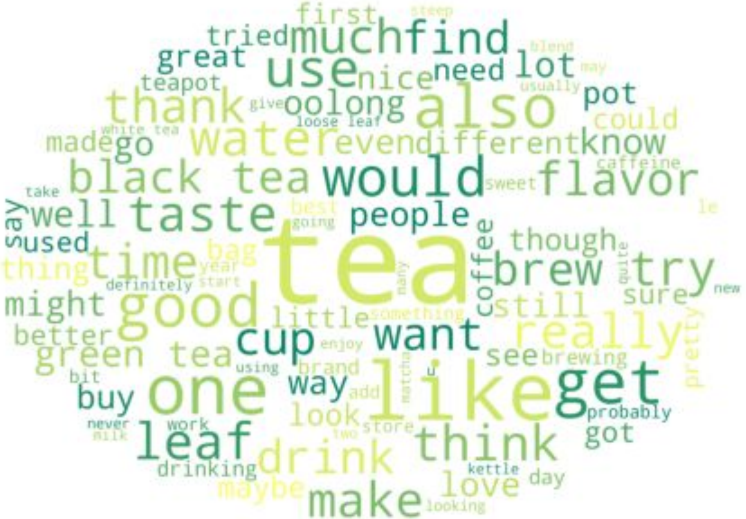


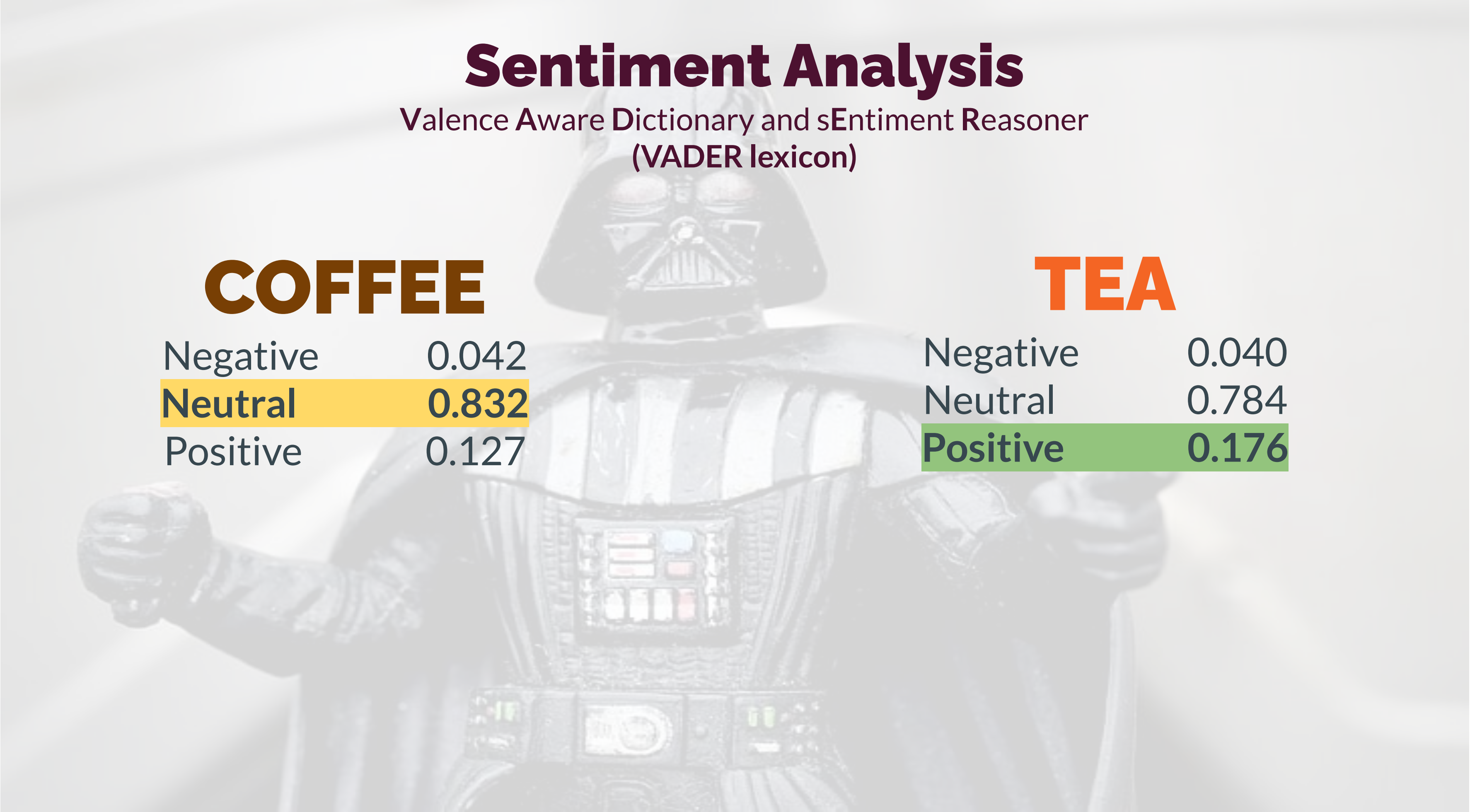
🟧 Coffee

🟦 Tea

Word Clouds


coffee
like
get
one
grinder
water
also
would
use
brew
bean
really
need
make
time
much
think
good
espresso
roaster
different
taste
know
using
could
still
might
grind
roast
used
thing
flavor
le
two
maybe
may
say
u
question
great
pour
going
filter
cup
fine
price
machine
method
brewing
way
made
probably
mean
note
since
go
find
recipe
work
new
start
making
drink
take
bag
always
little
many
french press
part
got
people
sure
try
something
want
getting
looking
place
around
well
right
day
even
first
bit
aeropress
give
actually
see
help
best
year
lot
thank
cafe
buy
better
seem
ground


tea
like
one
good
get
would
black tea
taste
water
also
use
much
find
first
steep
tried
great
nice
need
lot
blend
may
teapot
usually
pot
thank
give
oolong
loose leaf
could
white tea
made
go
even
different
know
caffeine
sweet
flavor
take
well
best
going
people
le
say
used
though
coffee
thing
time
bag
year
start
definitely
many
brew
try
quite
new
sure
might
little
something
still
really
pretty
better
cup
want
green tea
enjoy
see
brewing
bit
using
brand
way
u
matcha
buy
add
look
store
never
milk
work
two
leaf
drink
think
probably
got
drinking
kettle
day
maybe
love
make
looking



Sentiment Analysis

Valence **A**ware **D**ictionary and s**E**ntiment **R**easoner

(VADER lexicon)

COFFEE

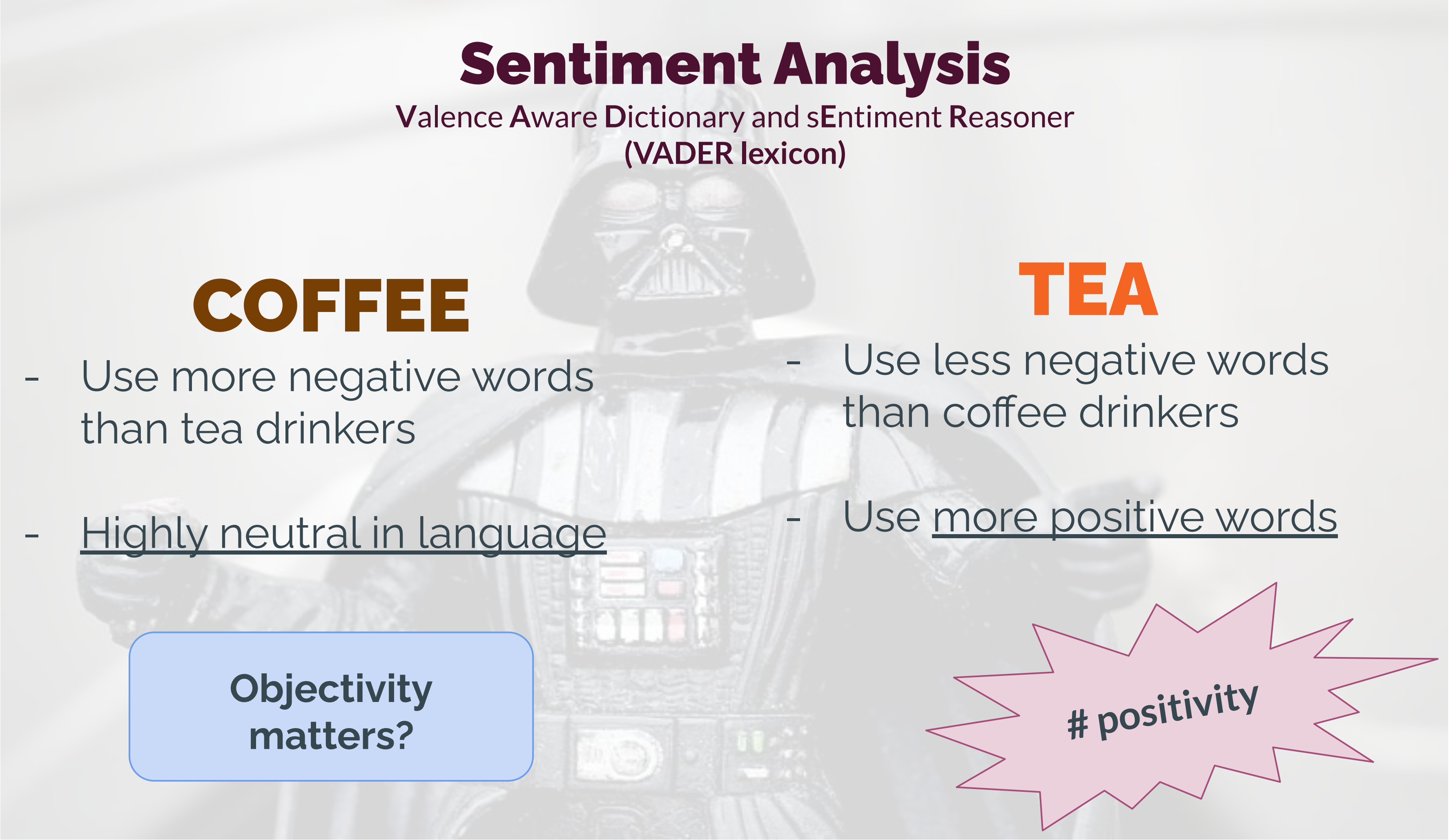
Negative	0.042
Neutral	**0.832**
Positive	0.127

TEA

Negative	0.040
Neutral	0.784
Positive	**0.176**

Sentiment Analysis

Valence Aware Dictionary and sEntiment Reasoner
(VADER lexicon)



COFFEE

- Use more negative words than tea drinkers
- Highly neutral in language

Objectivity matters?

TEA

- Use less negative words than coffee drinkers
- Use more positive words

# positivity

Classification Modelling

How did we predict if text comments were written by a coffee or tea drinker?

➔ Pre-processed text comments to extract key text features

➔ Separately employed combinations of “text vectorisation - supervised learning models” to obtain predictions

➔ Identified the best approach based on a set of performance metrics

Classification Modelling

(a) Pre-processed text comments to extract key text features

- **(1) Retained pre-apostrophe word sections of words with apostrophes [xx]**
- **(2) Removed stop words [xx]**
- **(3) Lemmatized words to retrieve their meaning [xx]**

Example: pre-processing of a comment from the r/coffee subreddit

Initial	i'm a morning coffee drinker but I never make my own at home … are there recommendations for travel mugs that will keep my coffee tasting good
After (1)	i a morning coffee drinker but I never make my own at home … are there recommendations for travel mugs that will keep my coffee tasting good
After (2)	morning coffee drinker never make home … recommendations travel mugs keep coffee tasting good
After (3)	morning coffee drinker never make home … recommendation travel mug keep coffee tasting good

Classification Modelling

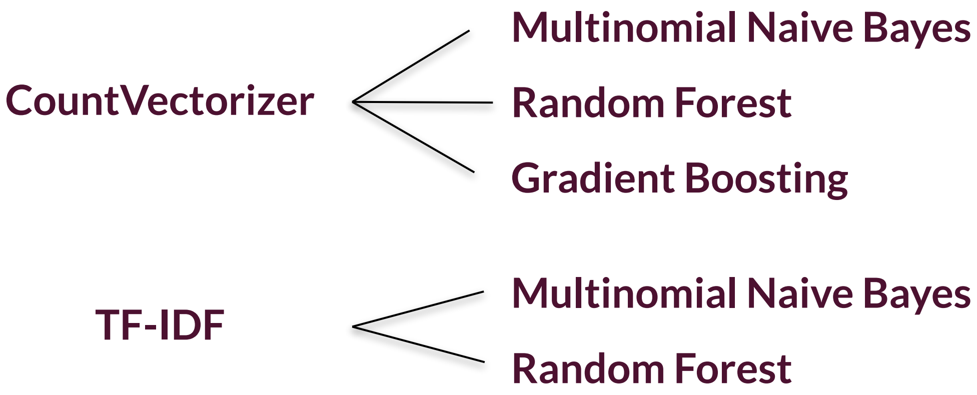
(b) Separately employed combinations of “text vectorisation - supervised learning models”

(c) Identified the best approach based on a set of performance metrics

Modelling Approaches Tested

Text Vectorizers

Supervised Learning Models



CountVectorizer

Multinomial Naive Bayes

Random Forest

Gradient Boosting

TF-IDF

Multinomial Naive Bayes

Random Forest

Performance Evaluation Metrics

Accuracy

% of predictions that are correct

F1-Scores

Measure that considers both

Precision : % of predicted positives that are true

Recall : % of actual positives predicted correctly

Performance Evaluation Metric Scores (1)

Model	M-NB	M-NB	RF	RF	G-Boost
Vectorizer	CountVec	TF-IDF	CountVec	TF-IDF	CountVec
Accuracy	**0.86**	0.71	0.84	0.70	0.85
F1: Coffee	**0.85**	0.72	0.81*	0.67	0.82*
F1: Tea	**0.86**	0.70	0.86*	0.72	0.86*

Performance Evaluation Metric Scores (2)

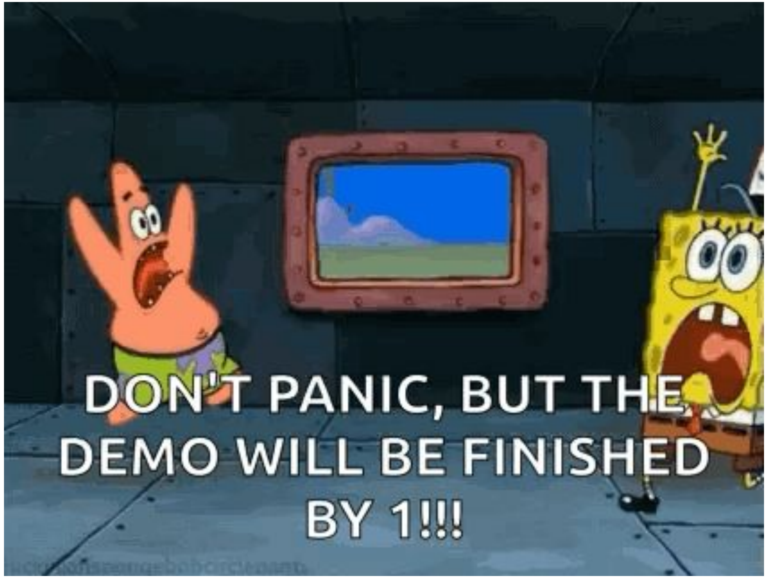
Model	**M-NB**	**M-NB**	**RF**	**RF**	**G-Boost**
Vectorizer	**CountVec**	**TF-IDF**	**CountVec**	**TF-IDF**	**CountVec**
F1: Coffee	**0.85**	0.72	0.81*	0.67	0.82*
• Precision	**0.88**	0.69	0.98	0.73	0.98
• Recall	**0.83**	0.76	0.69	0.63	0.70
F1: Tea	**0.86**	0.70	0.86*	0.72	0.86*
• Precision	**0.84**	0.74	0.76	0.68	0.77
• Recall	**0.89**	0.67	0.98	0.77	0.98

FINAL CHOSEN MODEL:

Multinomial Naive Bayes

(with Count Vectorizer)

DEMO



Key takeaways:

#1 Identification

We are able to provide …

tea / coffee drinker classification model that works reasonably well

#2 Continued Engagement

Tea: Images

Coffee: Descriptive words

#3 Continued Engagement

Tea: Types of tea leaves

Coffee: Equipment

Future Work:

(1) Train classification model on a wide range of text-based platforms

(2) Incorporate analysis of images