DACS CAFÉ

Brought to you by:
Desmond, Alvin, Clare, Soon Poh

ABOUT DACS

We are a group of data scientists looking to open a cafe by making *data-driven* business decisions.

PROBLEM STATEMENT

To analyze posts from the 'Coffee' and 'Tea' subreddits to develop a classification model capable of segregating texts into ‘Coffee’ and Tea’ categories

To use the sentiments gathered to:

- Pick out the most popular flavour profiles and drinks to create a data driven menu.
- Select the best coffee machine and coffee making tools for our cafe

SCOPE

01

Data Collection & Cleaning

02

Preprocessing & EDA

03

Modeling

04

Sentiment Analysis

05

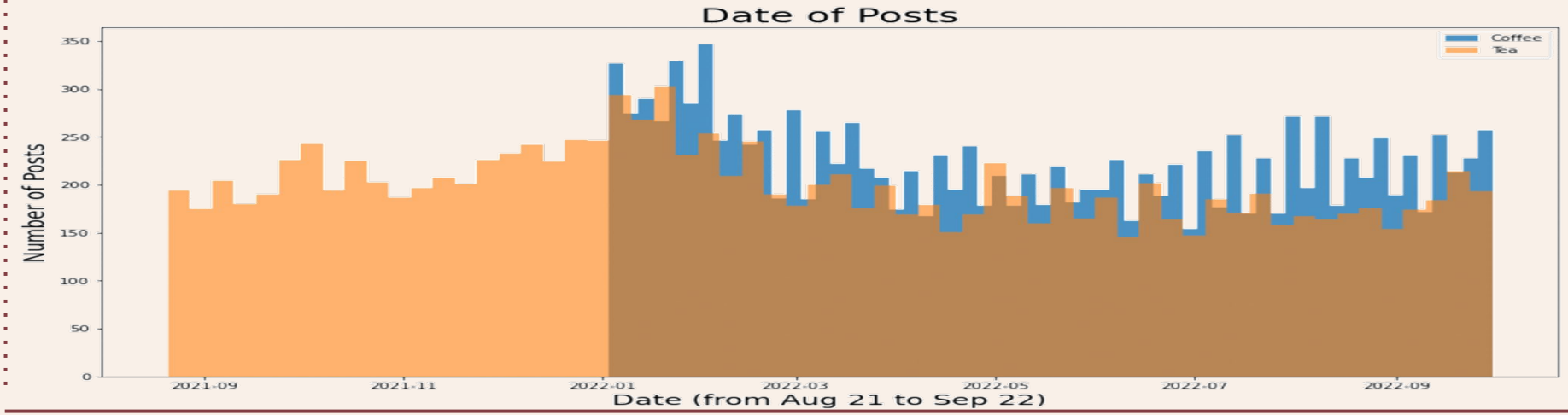
Insights & Recommendations

01

DATA COLLECTION & CLEANING

DATA COLLECTION AND CLEANING

Web Scraping.



DATA COLLECTION AND CLEANING

Verify authenticity of posts

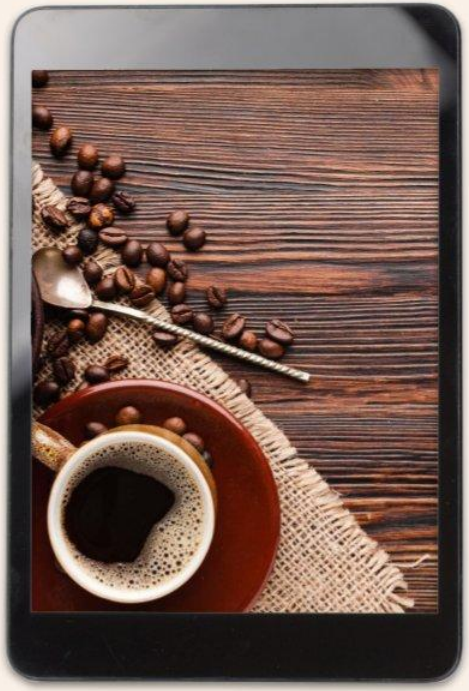
Drop duplicates / irrelevant posts

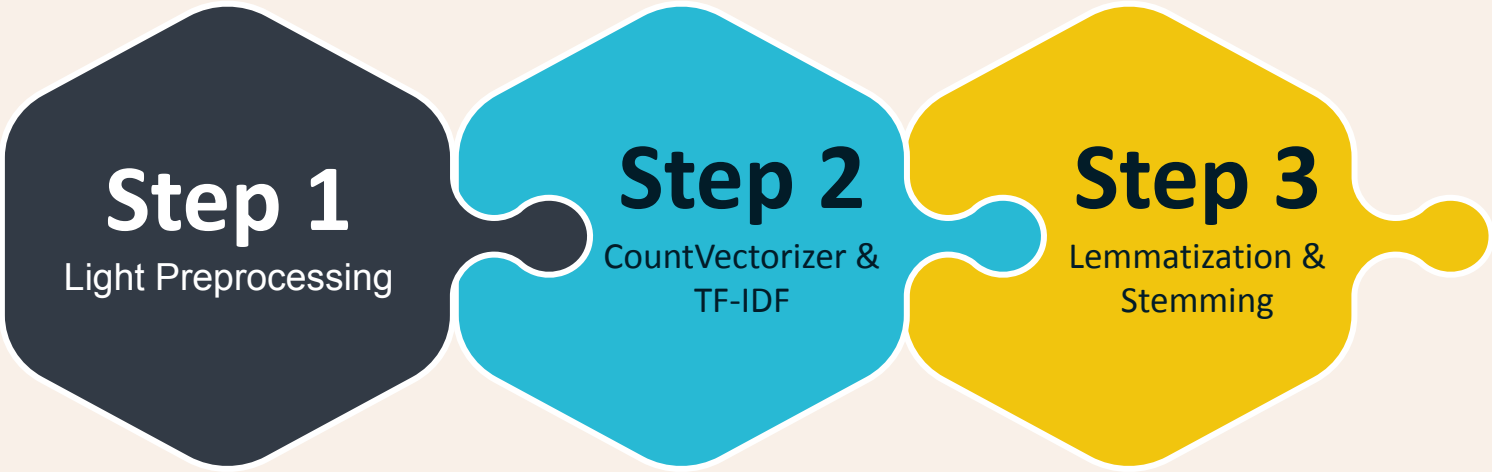
	subreddit	author	title	selftext	removed_by_category
4	tea	BigBart123	Sencha, Gyokuro, metals, etc.	[removed]	automod_filtered
5	tea	Haloreachyahoo	Optimal Tea Routine	[removed]	automod_filtered
11	tea	tajSL	What is the best brand of black tea in Sri Lanka?	[removed]	automod_filtered
17	tea	Royal_Ad_9082	I'm a bit obsessed	NaN	automod_filtered
25	tea	Eason-T	A pot of good tea 😊	NaN	automod_filtered

Create text and response Features

02

PREPROCESSING & EDA



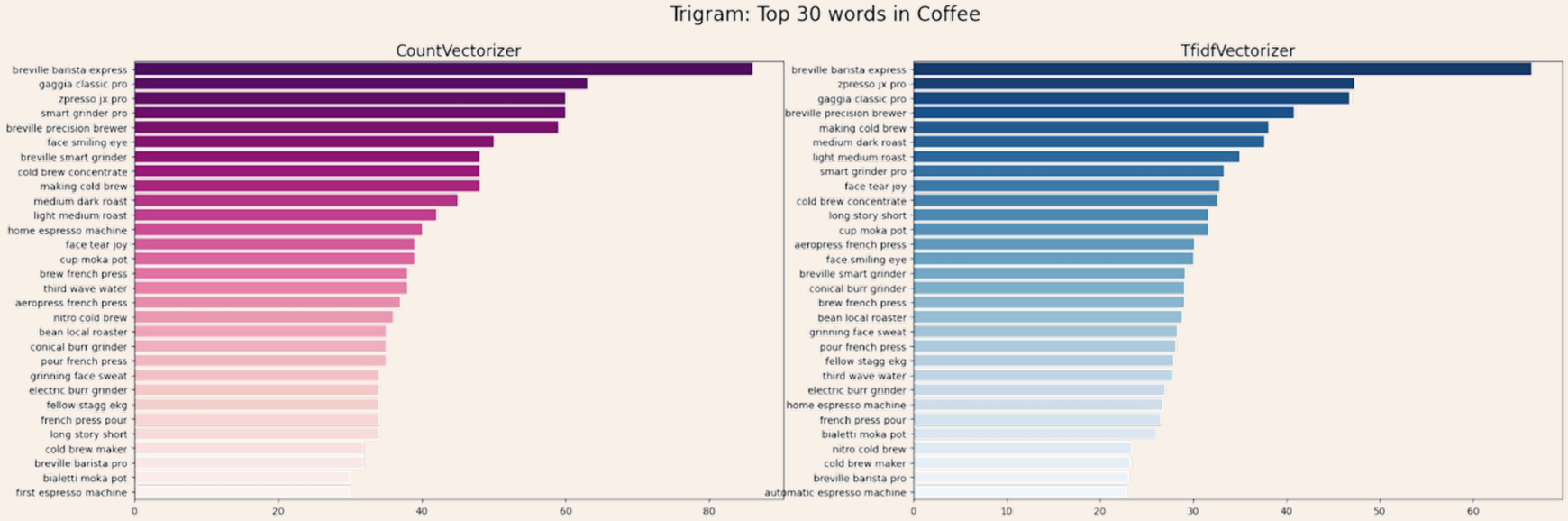


- Removal of http links
- Lowercase
- Removal of emojis
- Punctuations

- Dropped stop words, common words but kept key labels
- Identification of words of interest

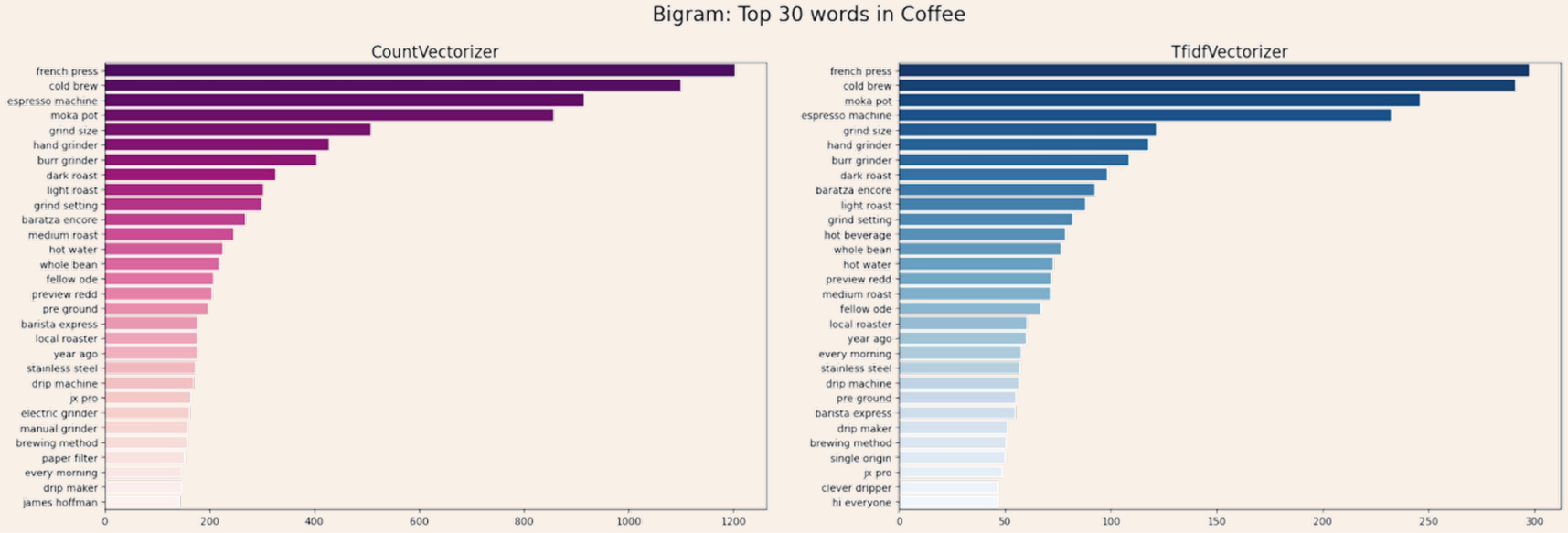
- Dropped key labels and customized words
- Comparison for number of features generated

WORDS OF INTEREST



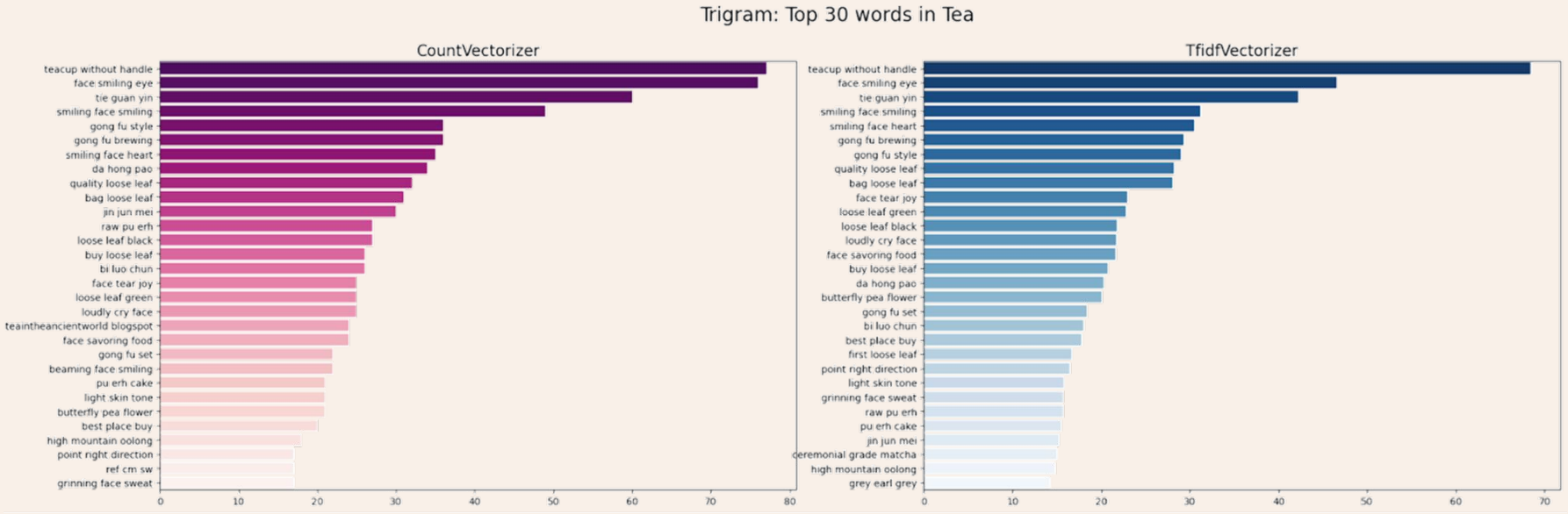
nitro cold brew, espresso, filter coffee, dark roast, burr grinder, timemore chestnut c2, breville barista pro, aeropress french press, gaggia classic pro

WORDS OF INTEREST



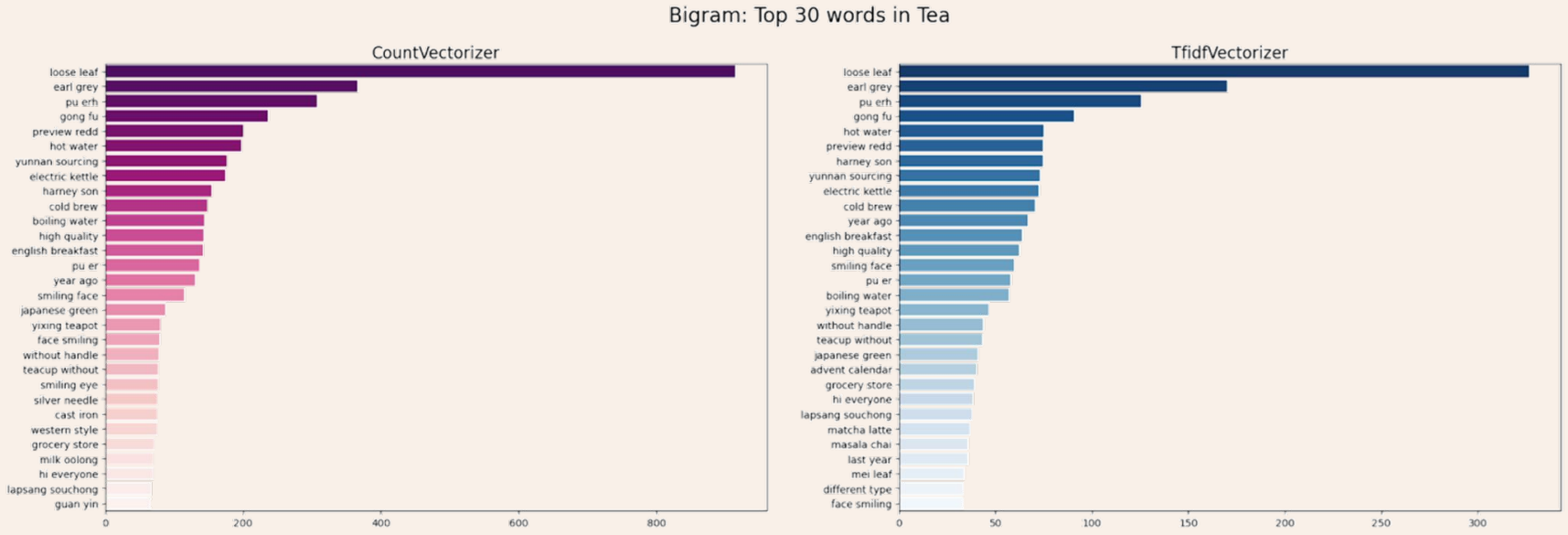
nitro cold brew, espresso, filter coffee, dark roast, burr grinder, timemore chestnut c2, breville barista pro, aeropress french press, gaggia classic pro

WORDS OF INTEREST



cold brew, jin jun mei, da hong pao, green tea, earl grey, japanese green tea, white tea, black tea, oolong tea, harney sons, english breakfast

WORDS OF INTEREST



cold brew, jin jun mei, da hong pao, green tea, earl grey, japanese green tea, white tea, black tea, oolong tea, harney sons, english breakfast

03

MODELING

PYCARET

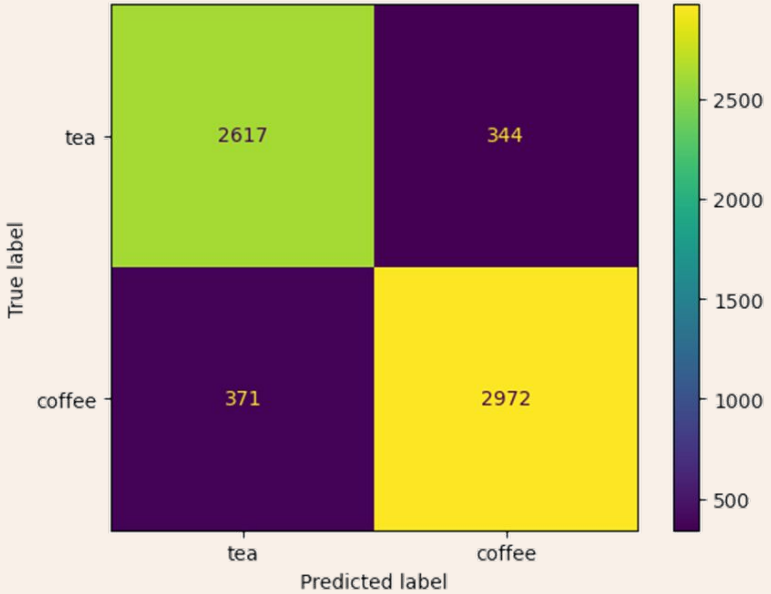
	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
lr	Logistic Regression	0.8820	0.9594	0.8826	0.8670	0.8746	0.7632	0.7635	1.1280
svm	SVM - Linear Kernel	0.8763	0.0000	0.8800	0.8587	0.8689	0.7519	0.7526	1.7020
et	Extra Trees Classifier	0.8733	0.9424	0.8588	0.8682	0.8634	0.7453	0.7454	12.0780
lightgbm	Light Gradient Boosting Machine	0.8689	0.9544	0.9104	0.8263	0.8662	0.7383	0.7418	2.2620
ridge	Ridge Classifier	0.8681	0.0000	0.8614	0.8567	0.8589	0.7351	0.7353	1.1680
rf	Random Forest Classifier	0.8655	0.9443	0.8719	0.8448	0.8580	0.7303	0.7308	6.2240
dt	Decision Tree Classifier	0.8335	0.8426	0.8320	0.8150	0.8233	0.6660	0.6663	10.9260
lda	Linear Discriminant Analysis	0.8298	0.9108	0.8338	0.8076	0.8204	0.6588	0.6593	36.6020
gbc	Gradient Boosting Classifier	0.8157	0.9211	0.9636	0.7287	0.8298	0.6372	0.6672	30.9720
nb	Naive Bayes	0.8149	0.8257	0.9285	0.7406	0.8239	0.6341	0.6523	0.5840
ada	Ada Boost Classifier	0.8106	0.9053	0.9491	0.7277	0.8237	0.6267	0.6531	7.3940
knn	K Neighbors Classifier	0.5948	0.6496	0.6506	0.5565	0.5995	0.1947	0.1974	63.8160
qda	Quadratic Discriminant Analysis	0.5415	0.0000	0.4758	0.3829	0.3638	0.0777	0.0969	67.1380

PYCARET

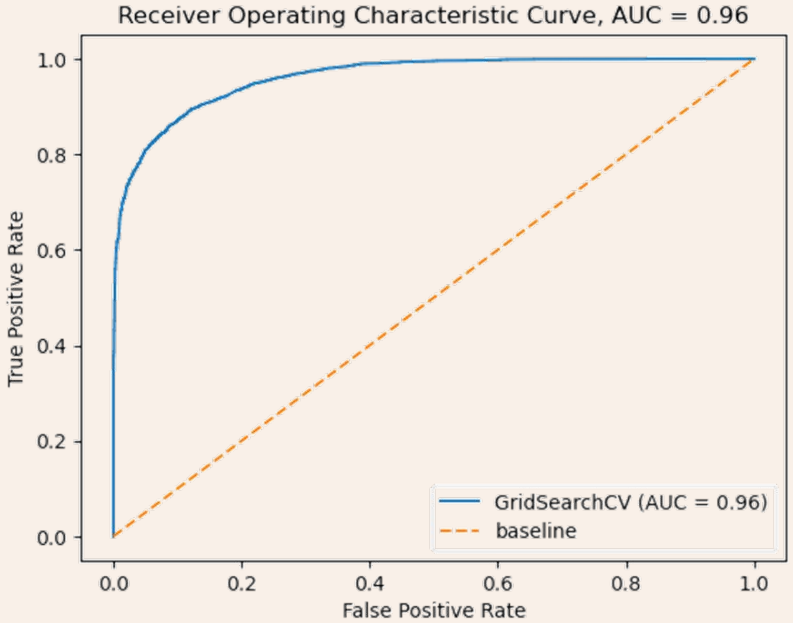
Model	Multinomial Naive Bayes (Baseline)	Logistic Regression	SVM - Linear kernel	Light Gradient Boosting Machine	Random Forest Classifier
Train score	0.8956	0.9207	0.9166	0.9067	0.9880
Test score	0.8715	0.8835	0.8803	0.8741	0.8735

- Logistic Regression best model before and after tuning
- Achieved the best score of 0.8835
- Accuracy as metric
- Utilised TF-IDF vectorizer

PYCARET



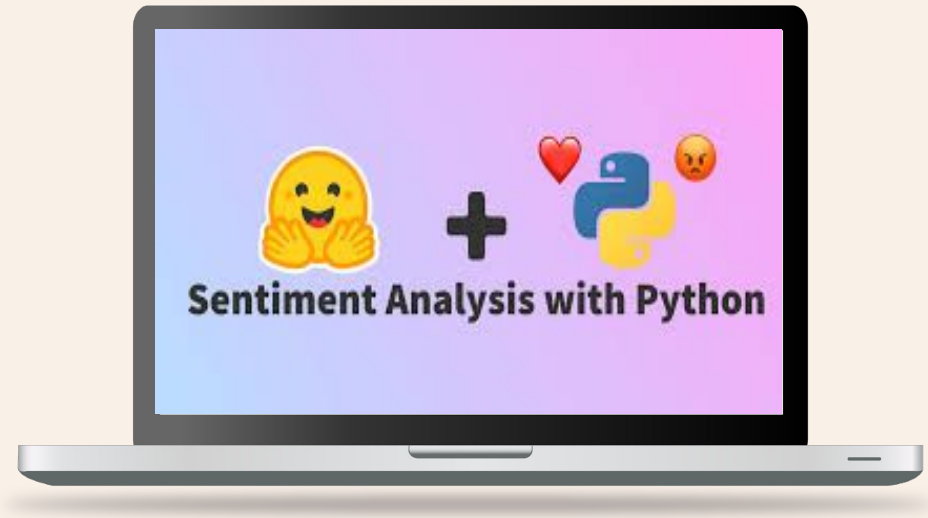
Confusion matrix



ROC curve

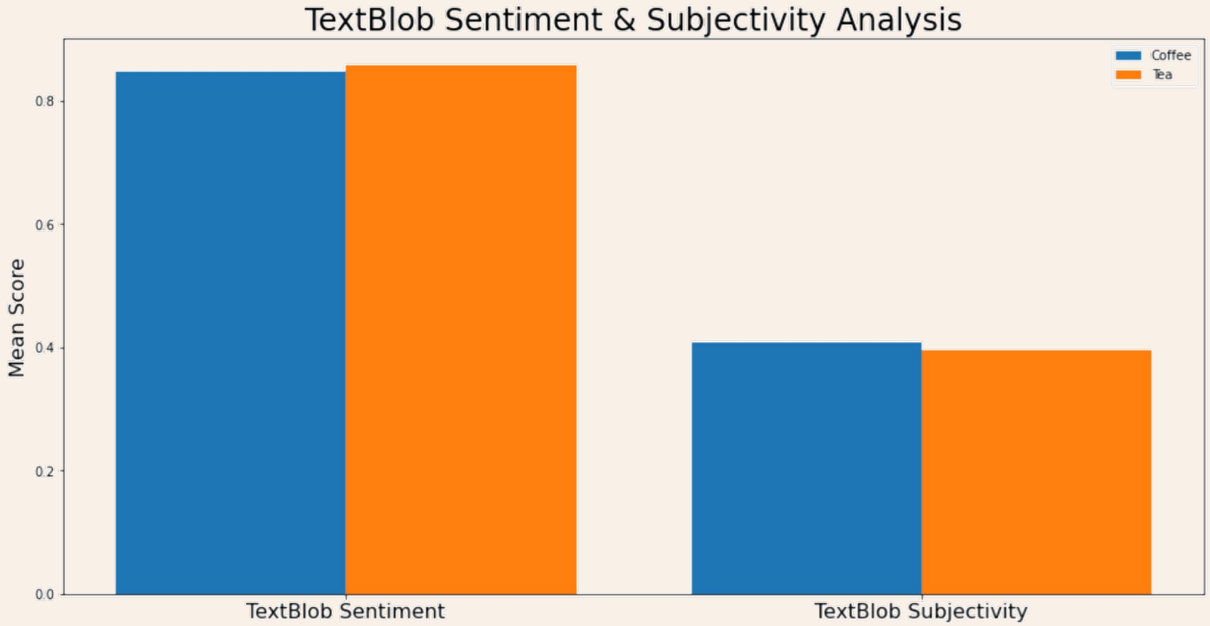
04

SENTIMENT ANALYSIS



General (Binary Baseline Model)

TextBlob



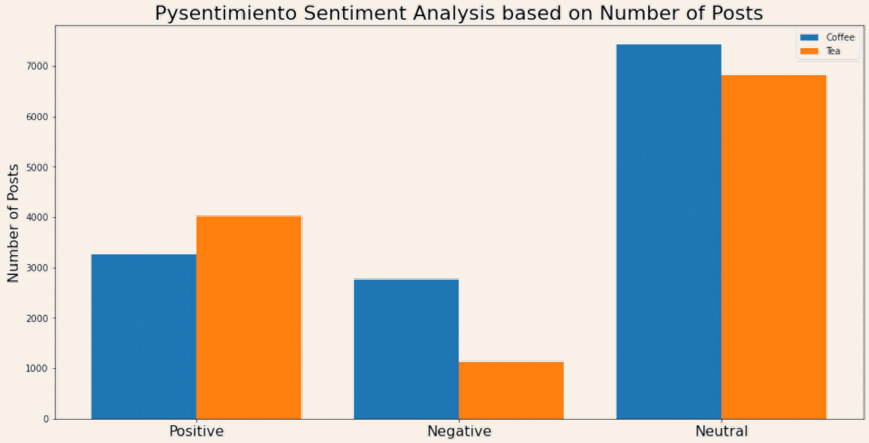
- Sentiment Score from -1.0 (-ve) to 1.0 (+ve)
- Classified data < 0 as 0, >0 as 1

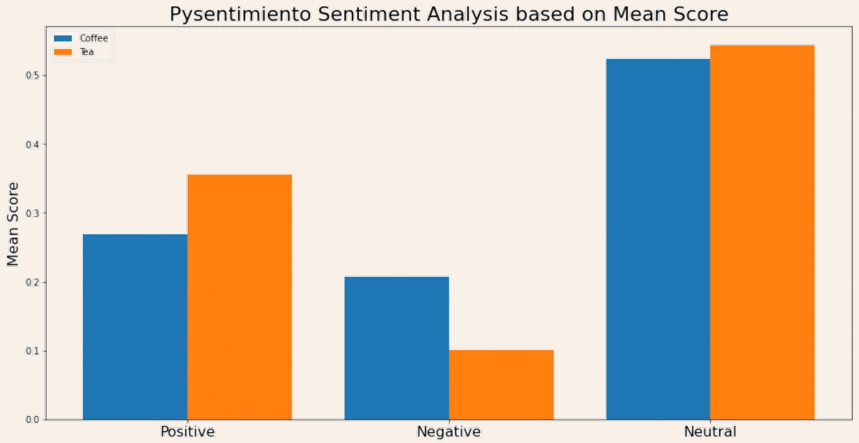
- Subjectivity Score from 0 (Objective) to 1.0 (Subjective)
- Classified data < 0.5 as 0, >= 0.5 as 1

Sentiment (Trinary Model)

pysentimiento

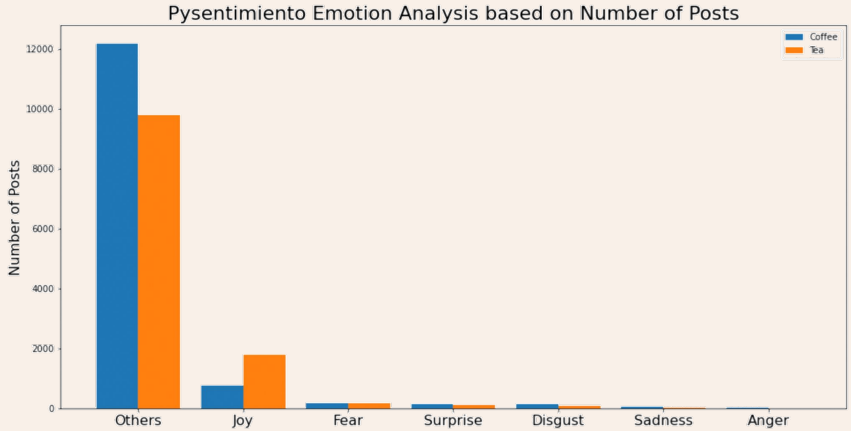
huggingface.co/pysentimiento



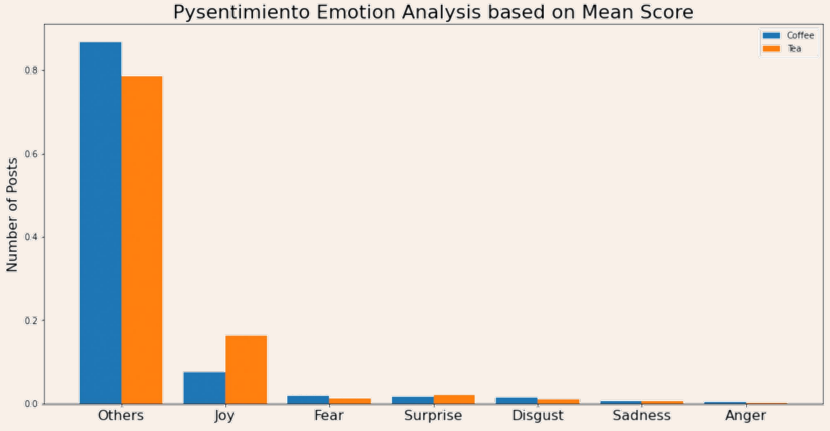


- Based on the highest score of the three sentiments (+ve, -ve, neutral) of each post
- Based on the mean score of each post

Emotion



- Based on the most predominant emotion of each post



- Based on the mean score of each post

Promising Topics

Coffee Subreddit

- Espresso
- Speciality Coffee
- Filter Coffee
- Dark Roast
- Baratza Encore
- Burr Grinder
- Drip Coffee
- Ground Coffee
- Moka Pot
- French Press
- Cold Brew
- Timemore Chestnut C2
- Breville Barista Pro
- Nitro Cold Brew
- Fellow Stagg Ekg
- 1Zpresso Jx Pro
- Aeropress French Press
- Breville Smart Grinder
- Breville Precision Brewer
- Gaggia Classic Pro
- Breville Barista Express

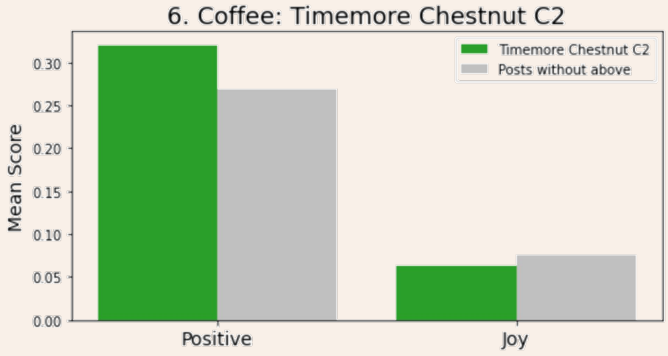
Tea Subreddit

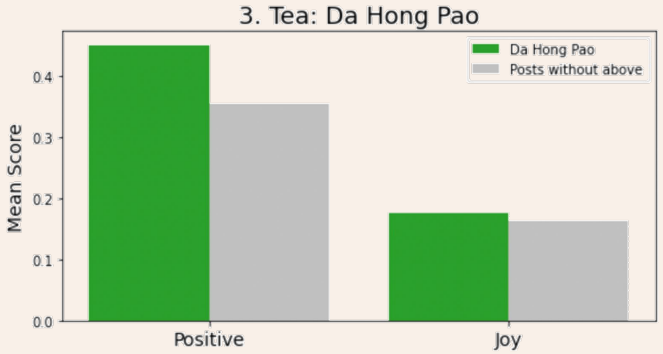
- Matcha
- English Breakfast
- Cold Brew
- Harney Sons
- Oolong Tea
- Green Tea
- Black Tea
- Gong Fu
- Earl Grey
- Da Hong Pao
- Tie Guan Yin
- Jin Jun Mei
- Bi Luo Chun
- Jasmine Green Tea
- Japanese Green Tea
- White Tea

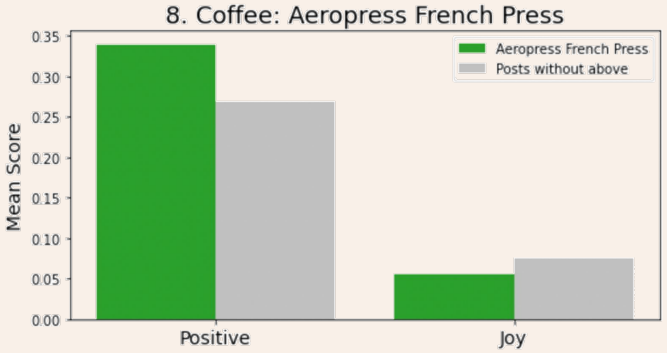
Best Performing Topics

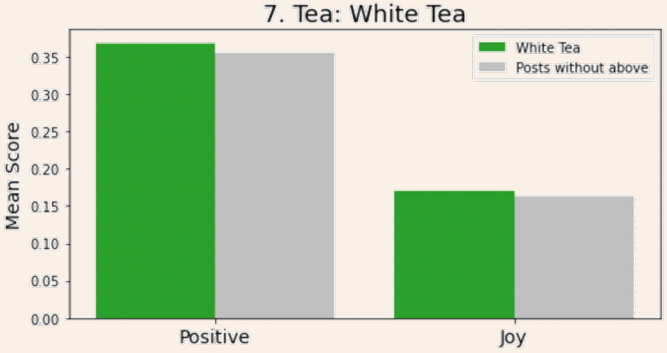
pysentimiento

huggingface.co/pysentimiento









Noteworthy Topics

Choice of Beverages

- Espresso
- Jin jun mei
- Da hong pao
- Green tea
- Earl grey
- Japanese green tea
- White tea
- Black tea
- Oolong tea
- English breakfast

Coffee/Tea-Making Method

- Nitro cold brew
- Filter coffee
- Dark roast
- Aeropress
- French press
- Cold brew tea

Products related to Coffee/Tea

- Burr grinder
- Timemore chestnut c2
- Breville barista pro
- Gaggia classic pro
- Harney and sons

05

CONCLUSION & RECOMMENDATIONS

Insights

- Coffee subreddit is more active than tea subreddit
- Less than 200 posts are related to coffee/tea brewing equipments (0.67% of dataset)
- Most Reddit users are from US
    - Almost 50% of desktop traffic on Reddit is from USA (Dec'21 -May '22)
- Logistic regression is the best model with 88% accuracy
- Through sentiment analysis, we are able to streamline and identify noteworthy topics in the areas of beverage menu, coffee/tea-making methods and coffee-making machines for cafe set-up

Limitations

- Logistic regression assumes that
  - Independent variables are linearly related to log odds
  - Absence of multicollinearity
  - Independence of observations
- Insufficient market data
  - The total number of appearances of top words make up a very low percentage of the dataset
- Time & Technical Constraints

Recommendations

- Collect more data to improve accuracy of model
  - Increase number of reddit submissions scrapped
  - Expand of subreddits to cafes as well to broaden our dataset
  - Explore more platforms such as quora, facebook, twitter etc
- Further customize stopwords list
- Sentiment analysis
  - Test other models to see which works best
  - Human annotated data
- Expand the model to other relevant topics to a cafe setup (Cafe design, food)
- Explore Gensim to gather more topics for analysis

THANKS

DO YOU HAVE ANY
QUESTIONS?