Subreddit Classifier

r/coffee or r/tea?

Group 1: Adi, Amira, Joel, Joshua, Yong Lim

DSI-28-SG
27 May 2022

Agenda

01

Intro: Background & Problem Statement

02

Data Cleaning

03

Exploratory Data Analysis

04

Modelling & Model Evaluation

05

Conclusion & Next Steps

01

Introduction

Background

Problem Statement

Our company is launching an e-commerce platform, and we would like to build a **classification model** to accurately classify textual data into ‘coffee’ and ‘tea’.

Our web development team, can **optimise our recommenders system**, we would like to accurately suggest products and tailor ads belonging to the same class.

Our business insights team can also leverage on the classification model to **analyse customer feedback** received on social media platforms.

What makes a good classification model?

01

Highest accuracy score

Model is able to accurately classify 'coffee' as coffee and 'tea' as tea.

02

Minimal overfitting of data

Very small difference between train and test scores.

03

Clear distinction of important features

No overlap or ambiguity between the keywords identified for coffee and tea.

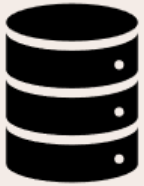
Scope of Data

reddit

>430m active users
Founded 23 June, 2005

r/coffee
Created: May 15, 2008
1M members

r/tea
Created: Dec 19, 2008
659k members



Data extracted:
1000 to 3000 posts from r/coffee
1000 to 3000 posts from r/tea
Posted between Mar 2022 and May 2022

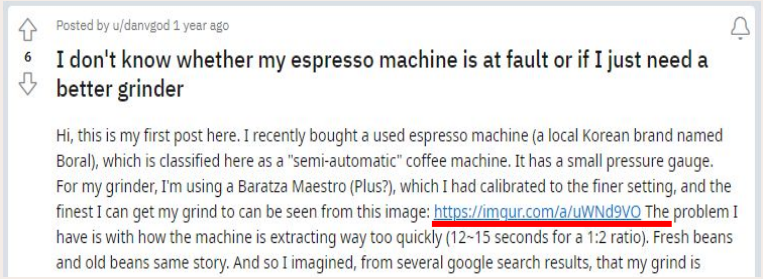
02

Data Cleaning

We replaced the following strings using ReGex

01

URLS



02

HTML terms e.g. '\n', '&#x200B'

03

- Punctuations
- Symbols
- Special characters
- Digits

04

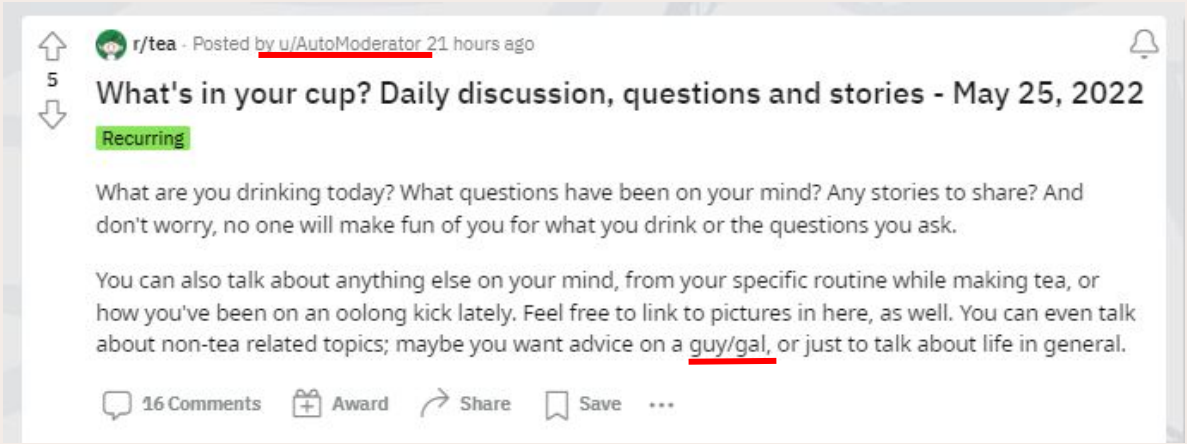
Removed duplicates based on title + self_text

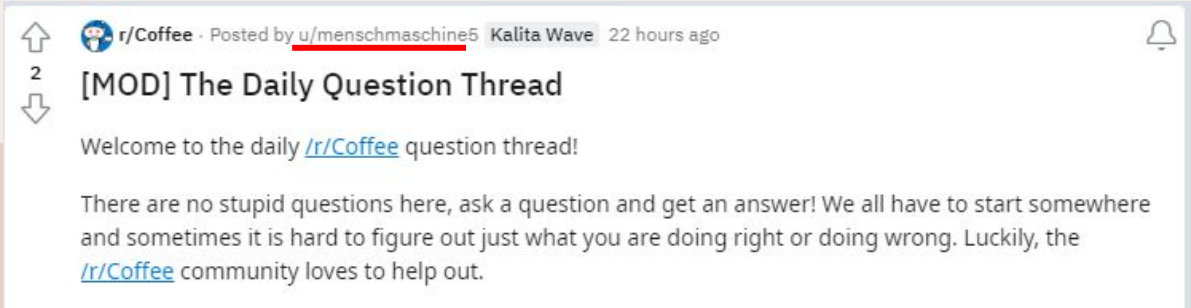
We filtered out [removed] or [deleted] posts

Possible Methods:

1. Filter out using ‘removed_by_category’
2. Filter out using ‘is_robot_indexable’

Several daily series of posts are creating a lot of noise in our data





Possible Methods:

1. Adjust **data extraction parameters**:

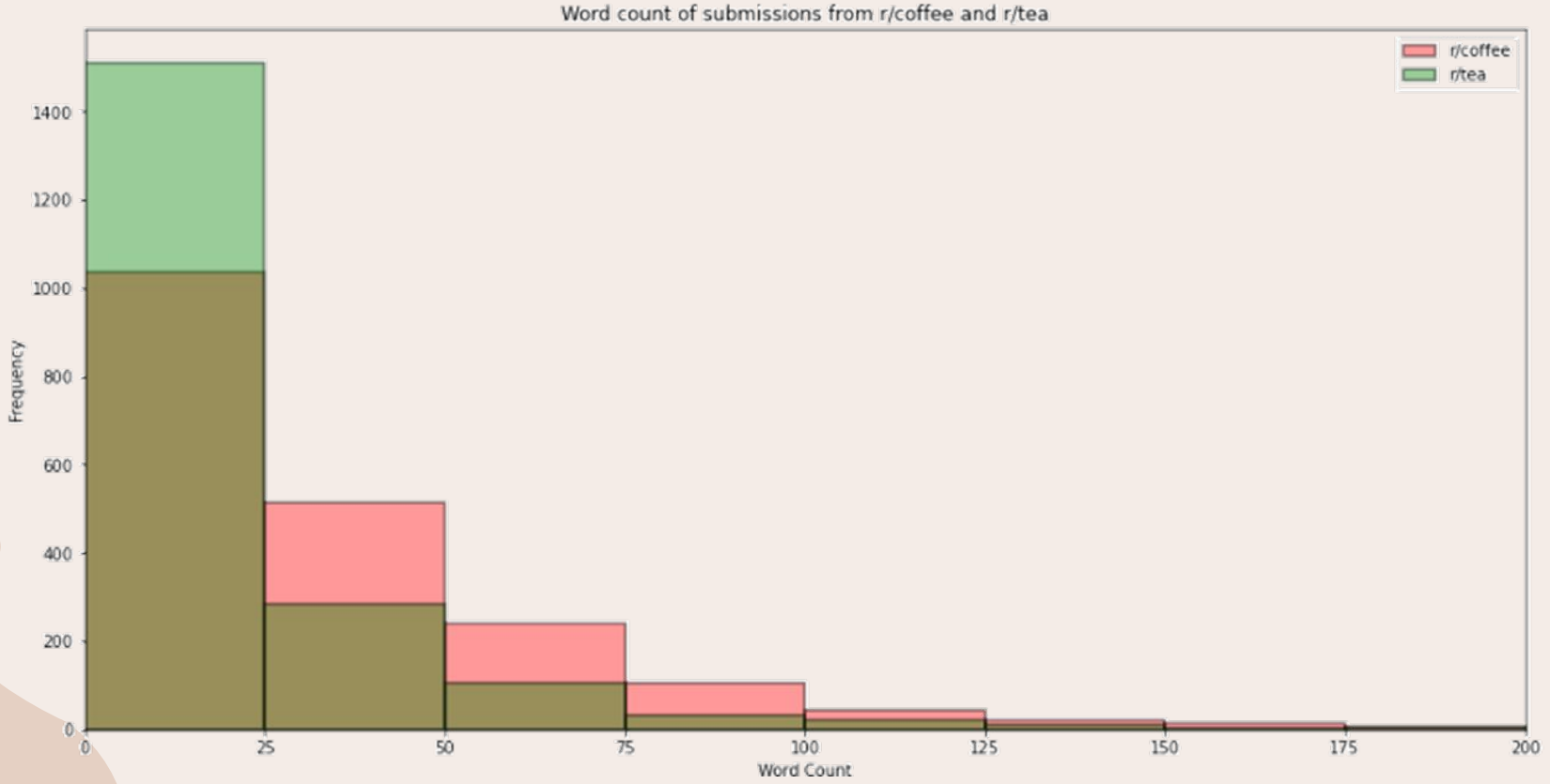
```
params={
    "subreddit": subreddit,
    "size": 100,
    "before": current_time,
    "stickied": False
}
```

2. Filter out posts based on **keywords** or **author**

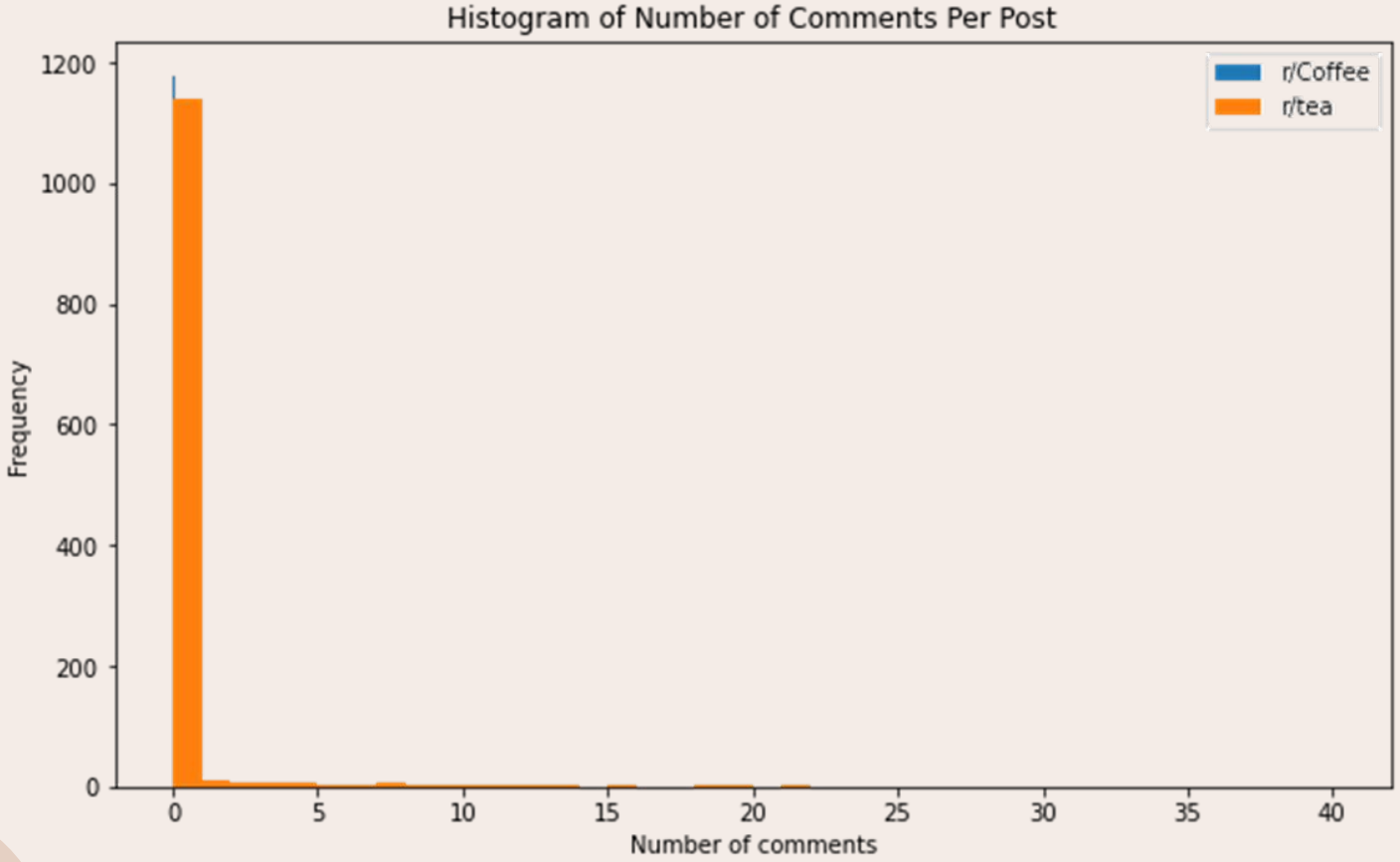
03

Exploratory Data Analysis

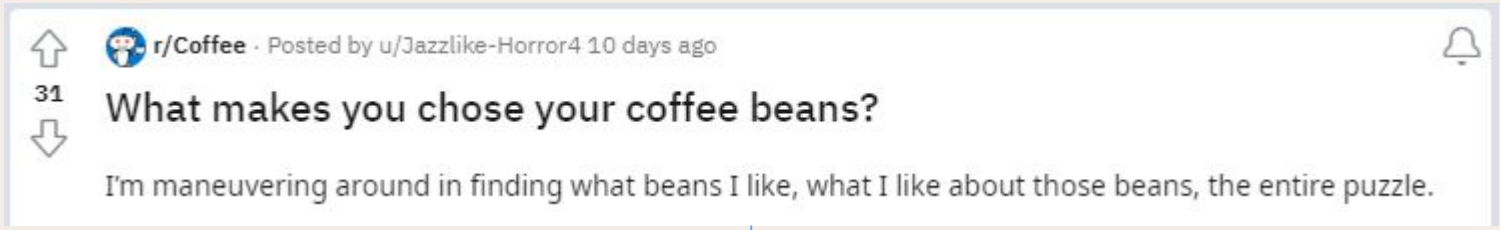
Submissions in r/tea tend to be a lot shorter than r/coffee



r/tea attracts more comments than r/coffee

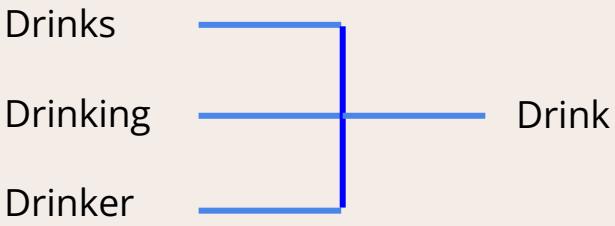


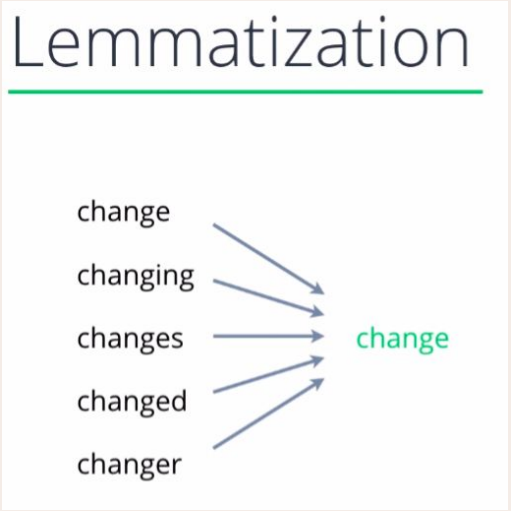
Tokenization



what makes you chose your coffee beans im maneuvering around in finding what beans i like what i like about those beans the entire puzzle

Lemmatization



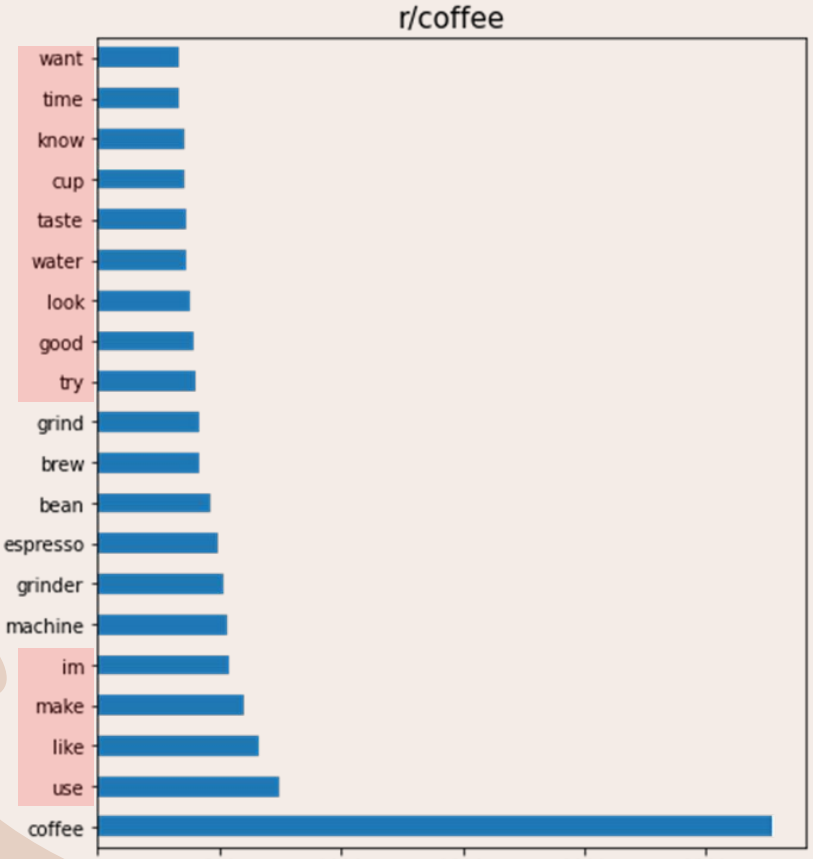


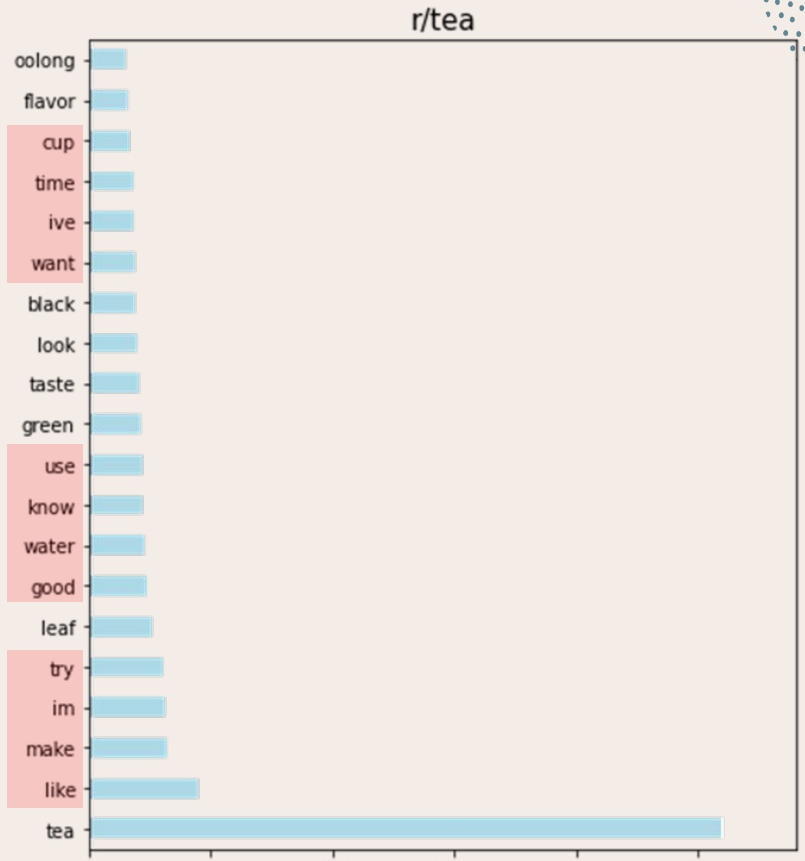
Natural Language ToolKit (NLTK) Stopwords

LIST OF ENGLISH STOPWORDS IN NLTK:

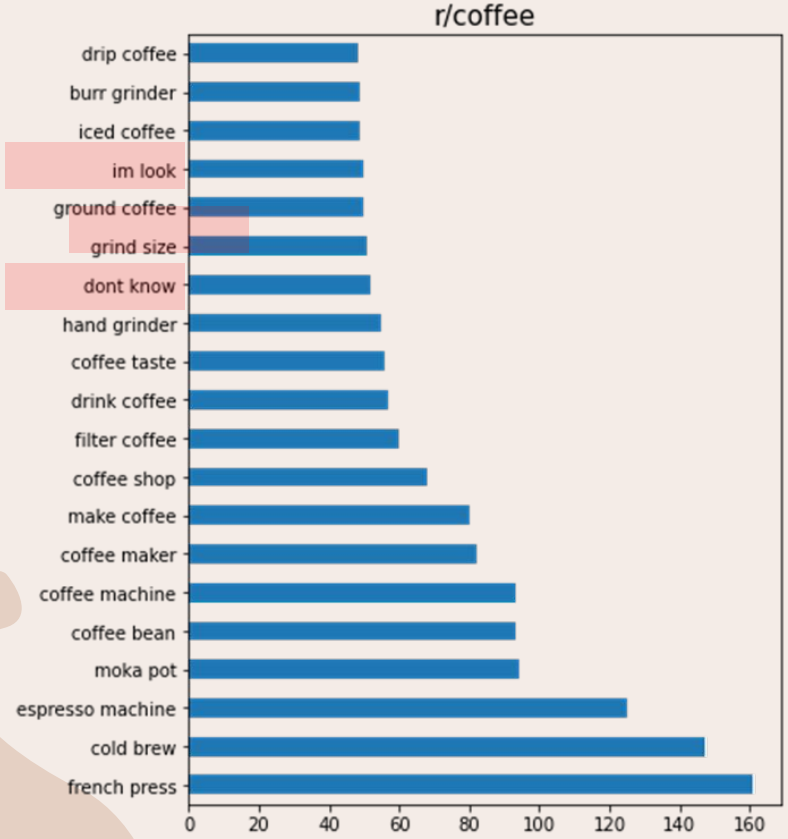
their, few, wasn't, has, m, or, did, isn, very, themselves, you've, you'd, do, between, other, t, shan, yourself, does, ours, i, it, should, what, himself, some, itself, there, weren, most, her, mustn, hers, doesn, won, doesn't, hasn, s, y, wouldn't, didn't, him, couldn, after, a, will, ain, than, for, being, which, during, ll, my, isn't, its, any, hadn't, his, then, don, of, shouldn't, out, our, have, such, o, nor, too, re, should've, needn't, same, she's, but, weren't, all, against, down, don't, can, you, under, where, wouldn, only, been, aren't, haven, that, doing, if, up, d, needn, ma, yours, shan't, wasn, because, about, those, he, are, was, at, hasn't, over, until, had, with, you're, below, haven't, mightn, here, own, off, both, whom, while, as, ourselves, they, further, mightn't, these, from, to, them, she, who, were, more, am, why, your, aren, hadn, in, won't, yourselves, no, me, didn, an, so, before, is, on, now, each, how, be, theirs, shouldn, mustn't, above, herself, just, you'll, the, through, again, once, having, by, when, myself, we, it's, this, that'll, couldn't, ve, and, into, not,

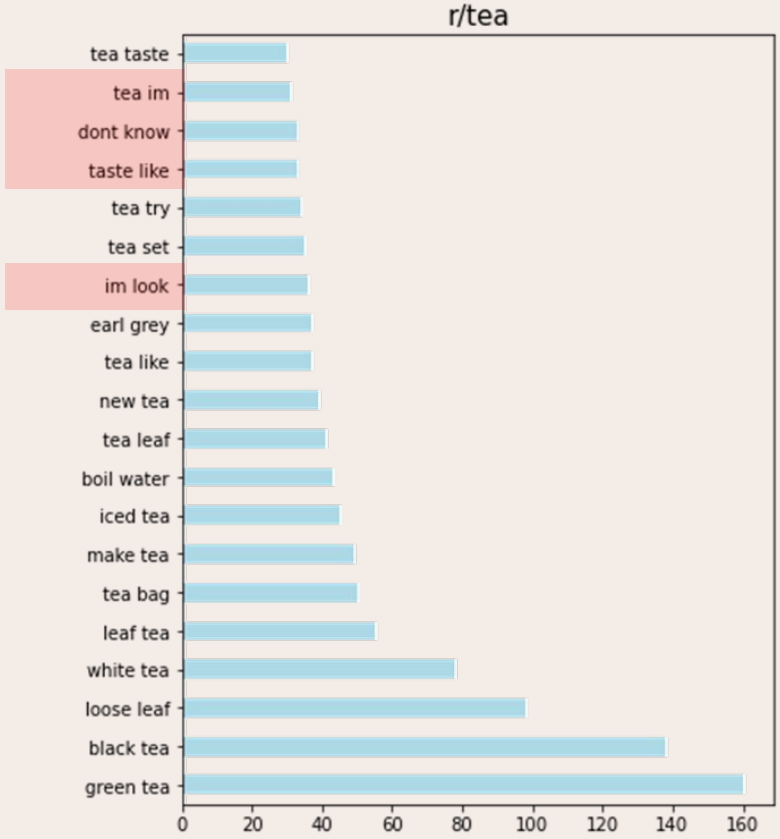
Top 20 words using standard stopwords



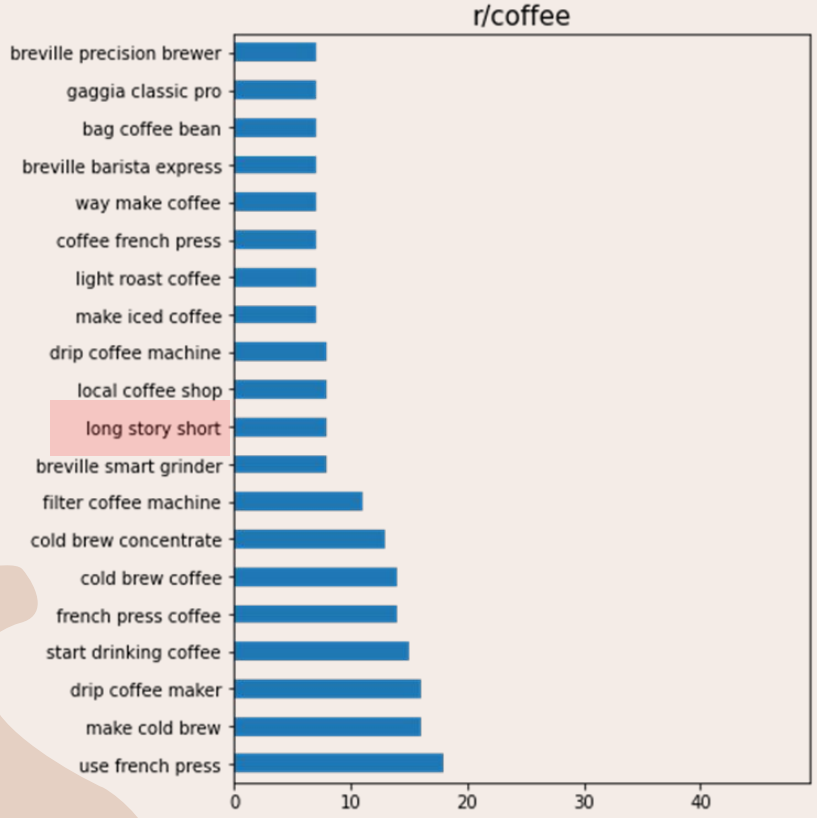


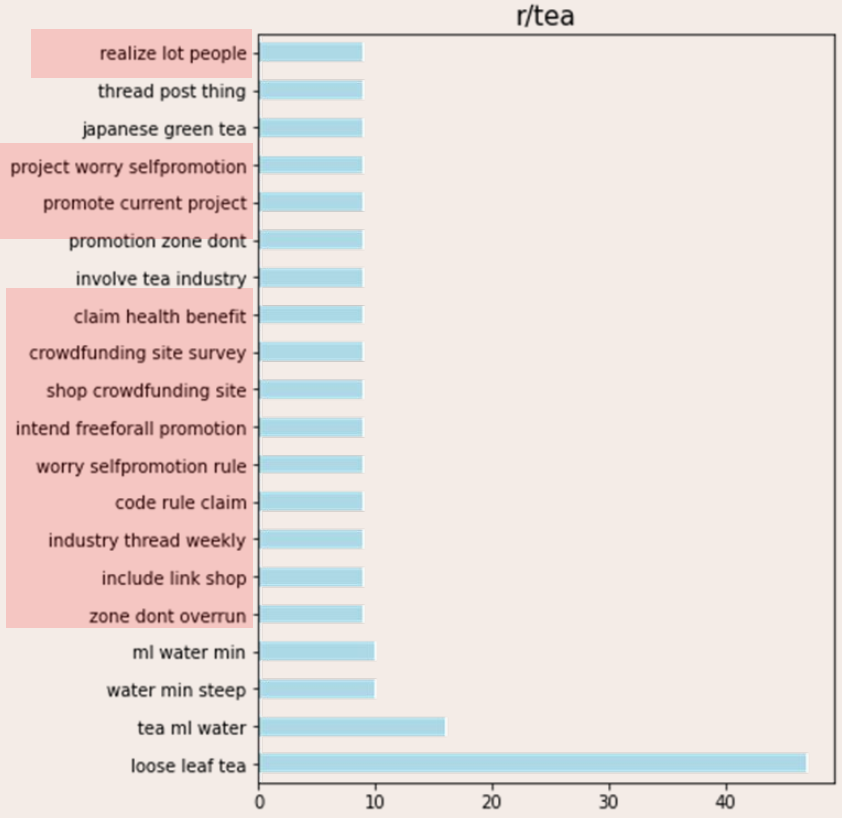
Top 20 two-word phrases using standard stopwords





Top 20 three-word phrases using standard stopwords

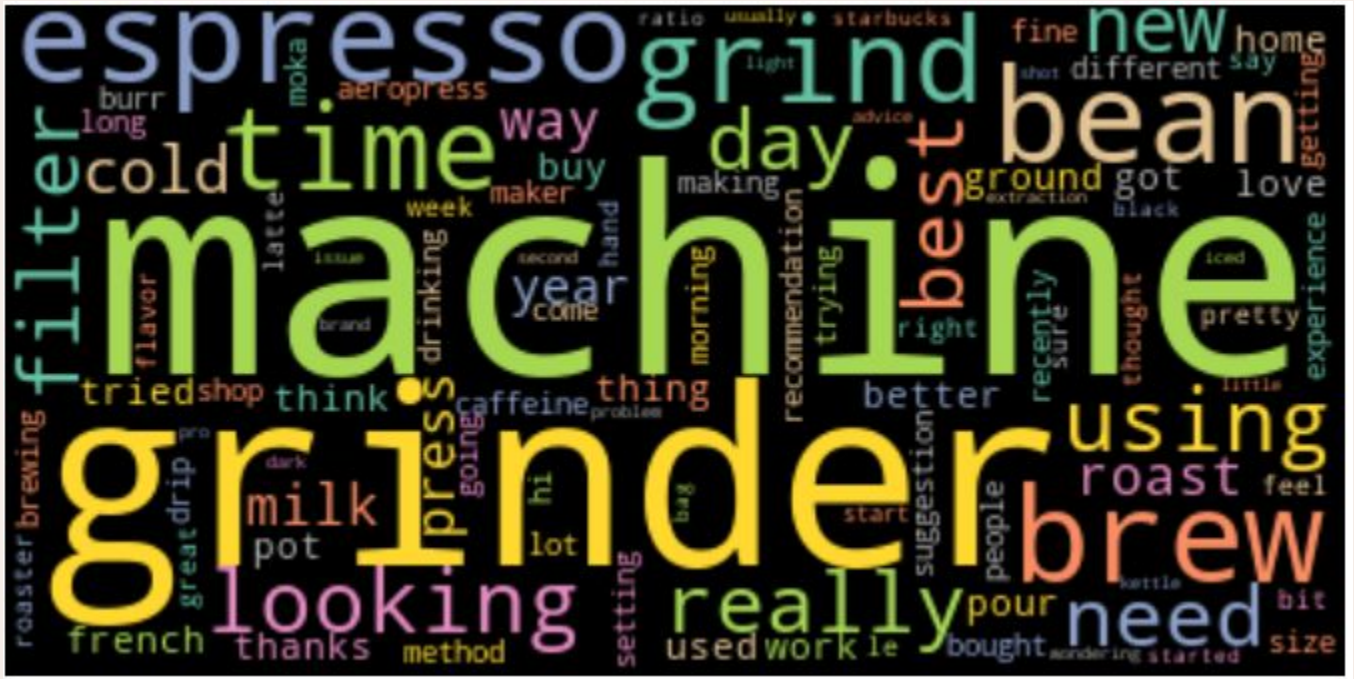




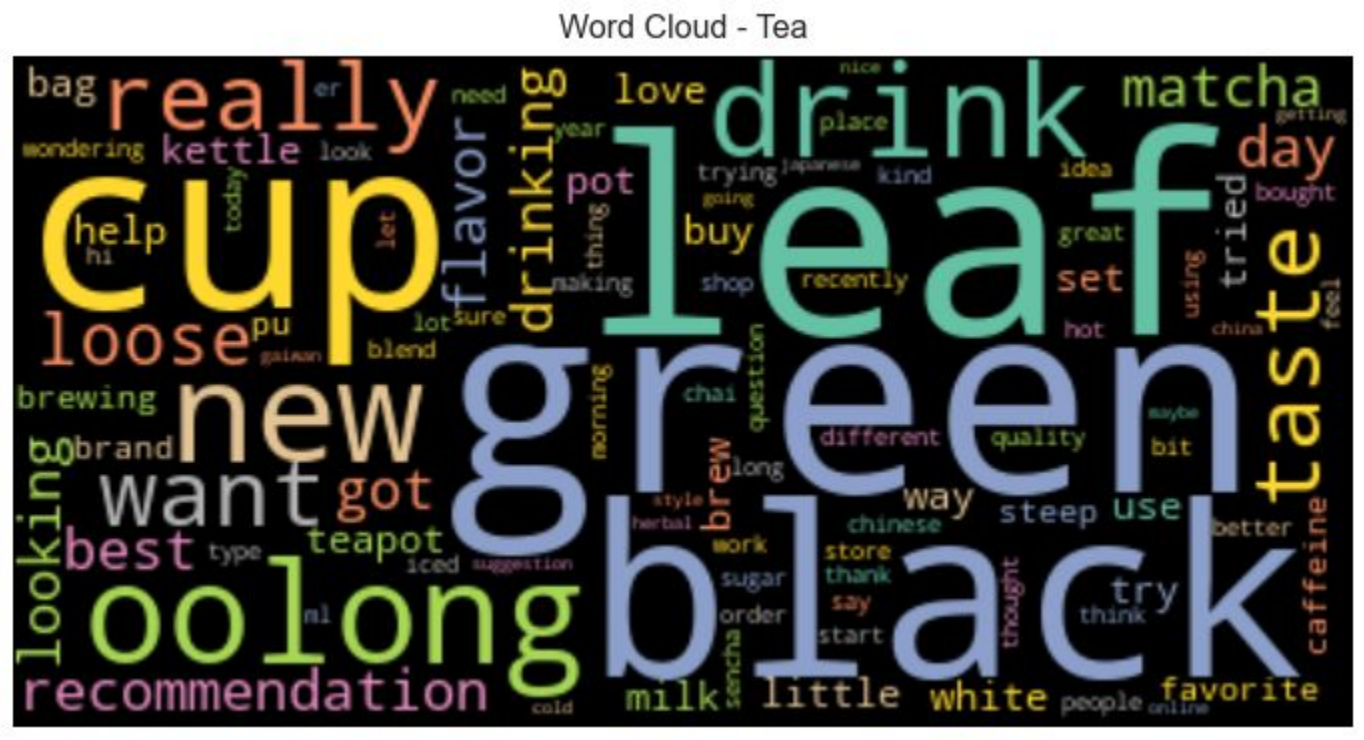
Custom stopwords

['tea','coffee','im','like','ive',
'try','water','taste','drink','know',
'want','cup','taste','look',
'question','use','dont',
'make','help','good']

Top words in r/coffee after applying custom stopwords



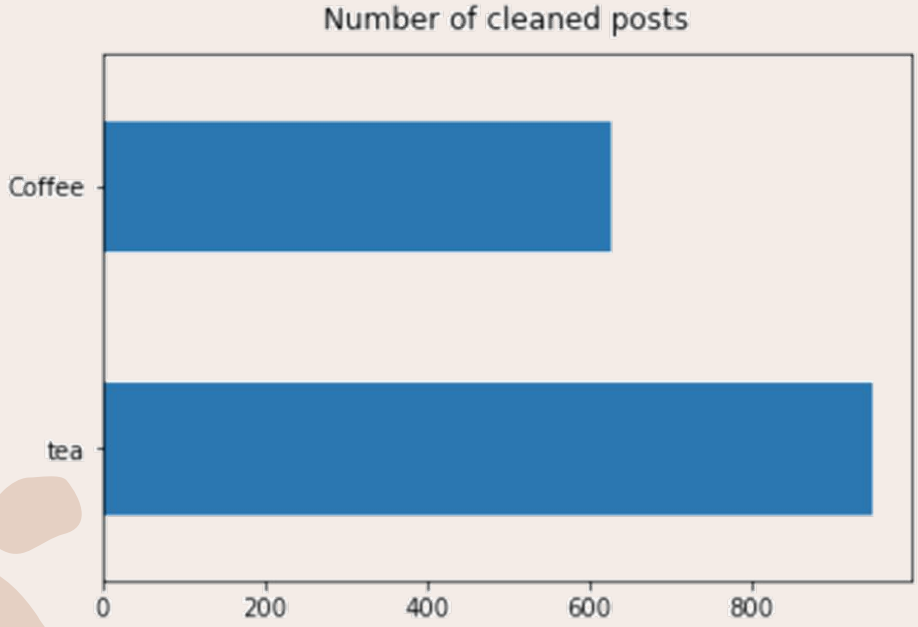
Top words in r/tea after applying custom stopwords



04

Modelling

Baseline Model Accuracy



Accuracy

$$\frac{\text{950 (Number of tea posts)}}{\text{1578 (Total posts)}} = \mathbf{0.602}$$

Transformer Model

Count Vectorizer

	the	red	dog	cat	eats	food
1. the red dog →	1	1	1	0	0	0
2. cat eats dog →	0	0	1	1	1	0
3. dog eats food →	0	0	1	0	1	1
4. red cat eats →	0	1	0	1	1	0

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

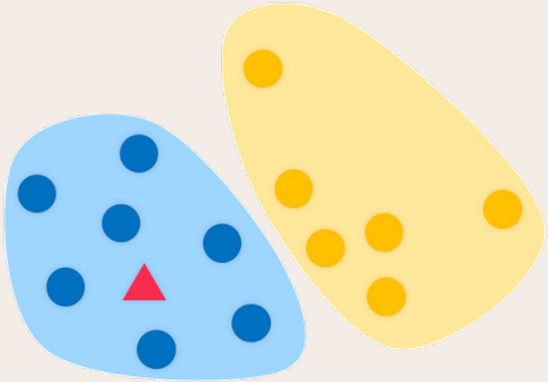
Term Frequency: Ratio of certain word found in a single post

Document Frequency: Ratio of number of posts that include certain word

Classifier Model

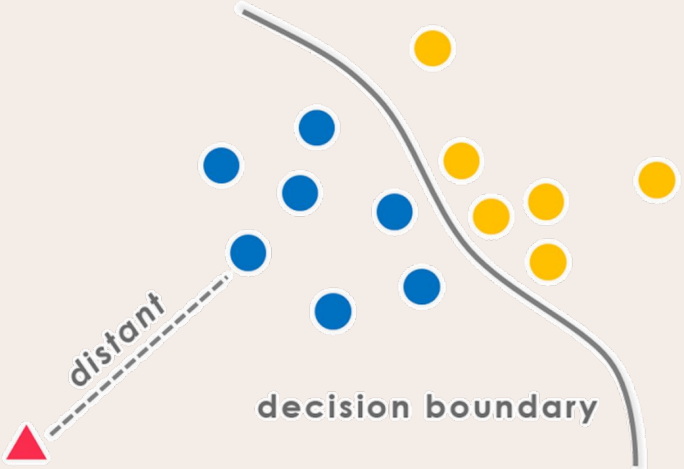
Multinomial Naive Bayes

Generative

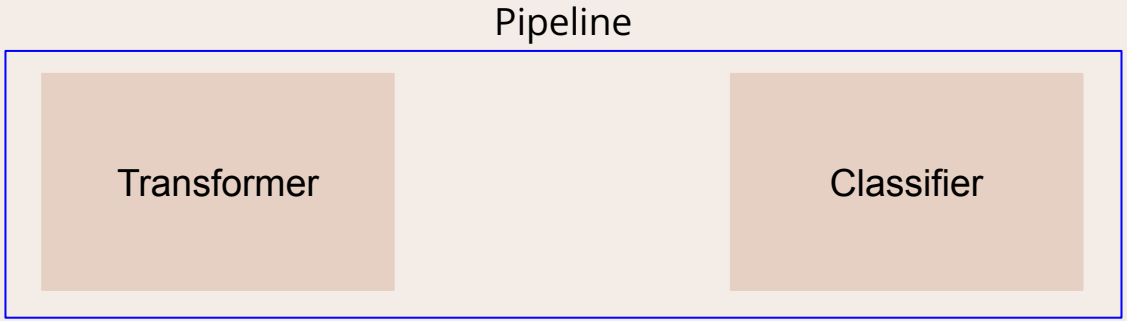


Logistic Regression

Discriminative

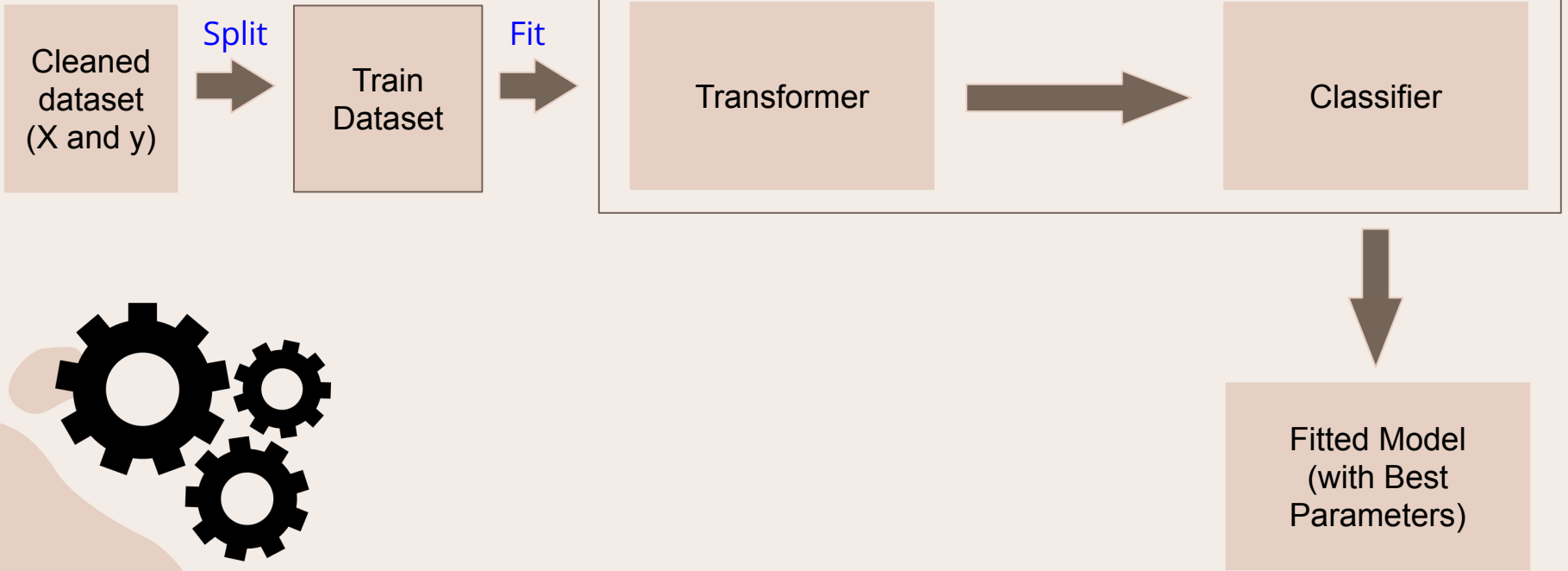


Modeling pipelines

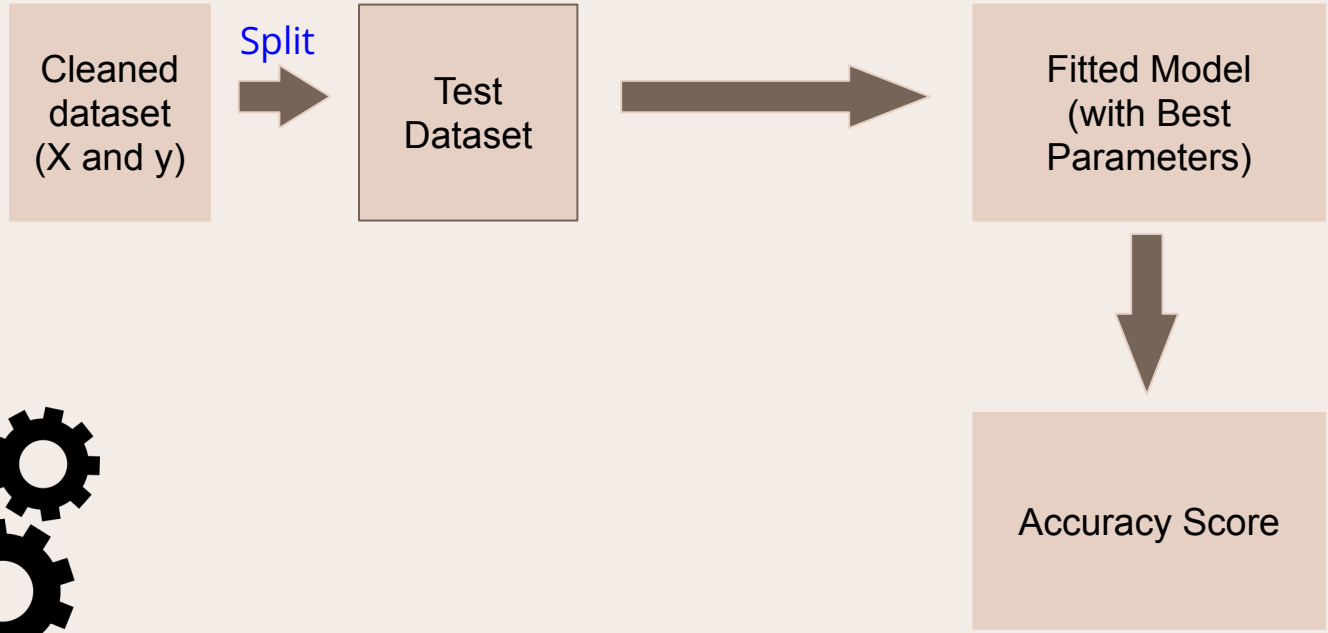


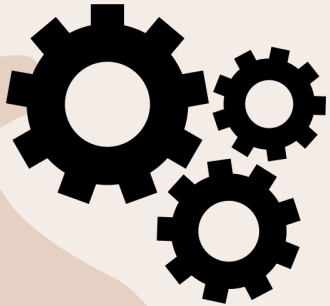
	Multinomial Naive Bayes	**Logistic Regression**
Count Vectorizer	Pipe 1	Pipe 2
TF-IDF Vectorizer	Pipe 3	Pipe 4

Hyperparameter Tuning (GridSearchCV)

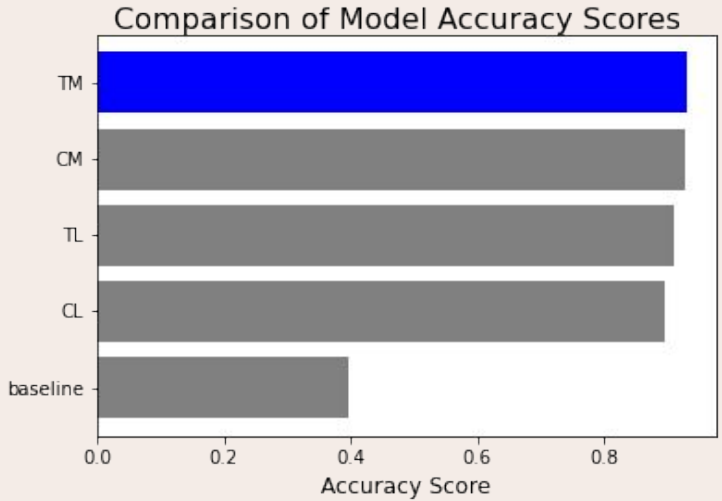


Obtain Accuracy Score





Best Model Pipeline: TF-IDF + Multinomial NB



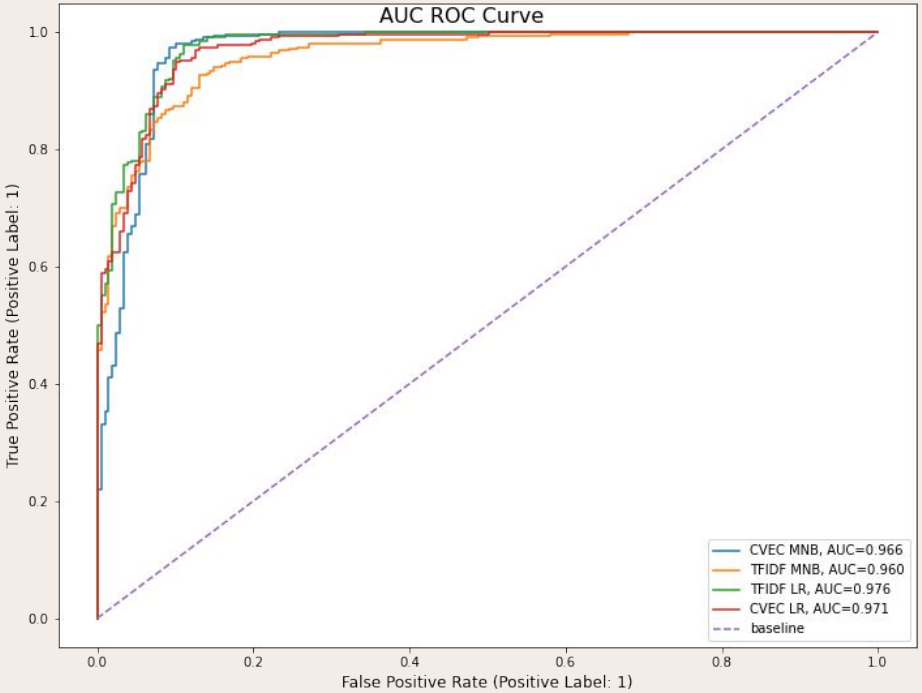
- **TM: Term Frequency-Inverse Document Frequency (TF-IDF) + Multinomial Naive Bayes (Accuracy = 0.93)**
- CM: CountVectorizer + Multinomial Naive Bayes
- CL : CountVectorizer + Logistic Regression
- TL: TF-IDF + Logistic Regression

Details of scoring

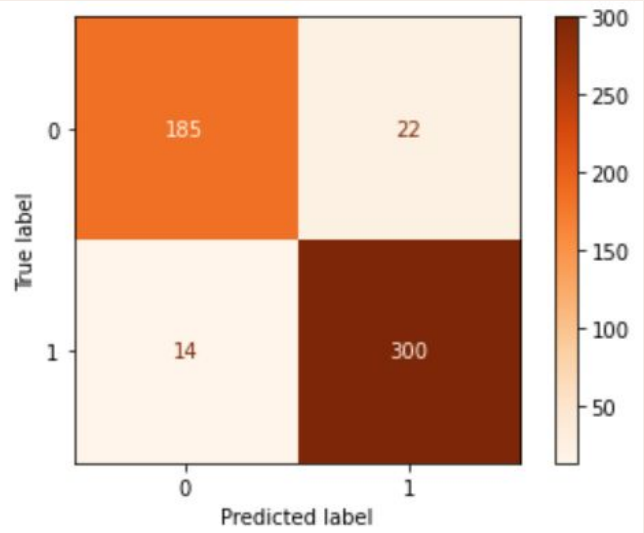
Accuracy	pipe_cm	pipe_cl	pipe_tm	pipe_tl
best score	0.928110	0.921497	0.934736	0.913914
training score	0.964995	0.992431	0.992431	0.955535
test score	0.927063	0.896353	0.930902	0.909789
ROC AUC	0.966353	0.959583	0.975692	0.970584

Good AUC ROC performance by all models

- All models exhibited a very high Receiver Operating Characteristic Curve Area Under Curve (AUC ROC)
- Tight range: 0.96 - 0.976 (the closer to 1, the better)



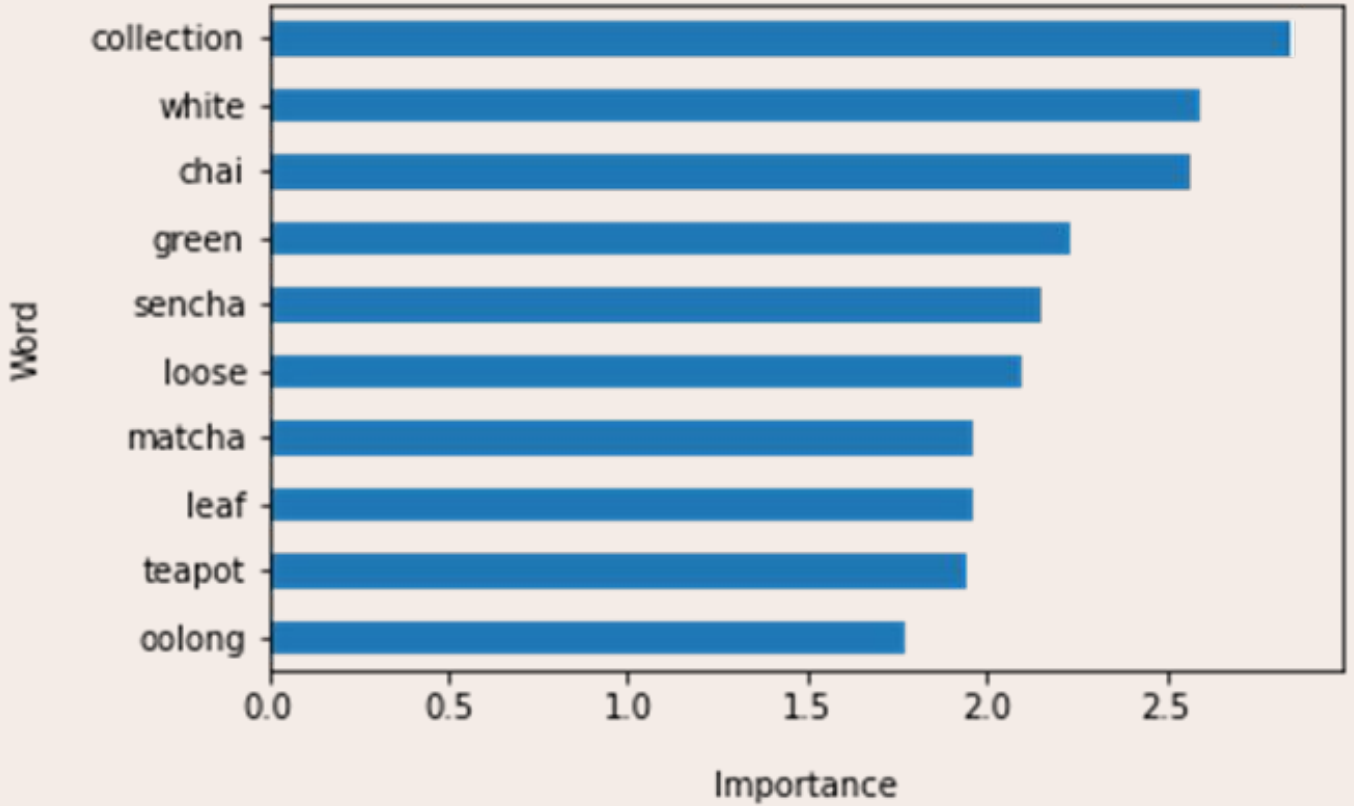
Good classification metric values for TF-IDF + Multinomial NB



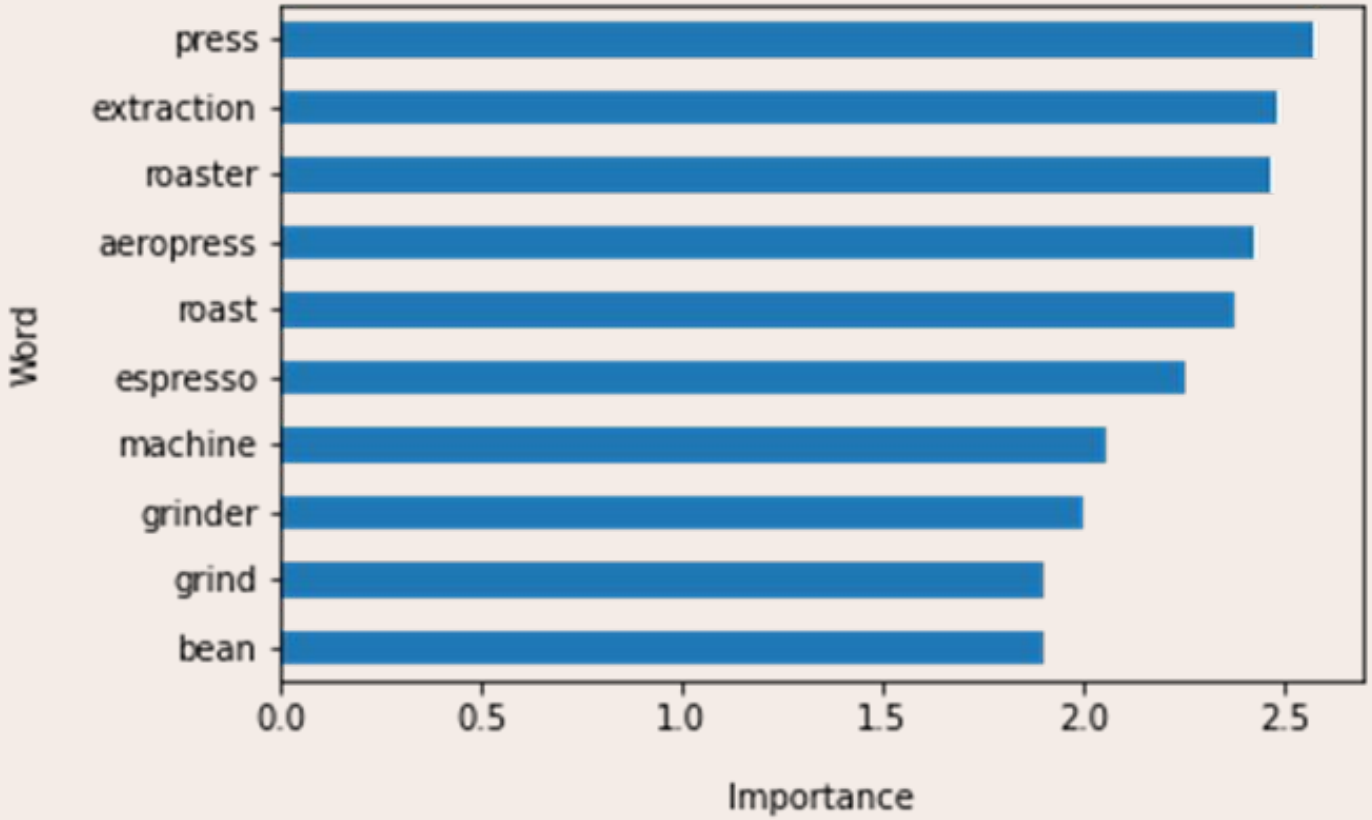
	Precision	Recall	F1-score	Support
Tea	0.93	0.96	0.94	314
Coffee	0.93	0.89	0.91	207

Our chosen model generally performs quite well in other evaluation metrics besides accuracy.

Top 10 Tea predictor words



Top 10 Coffee predictor words



05 Conclusion

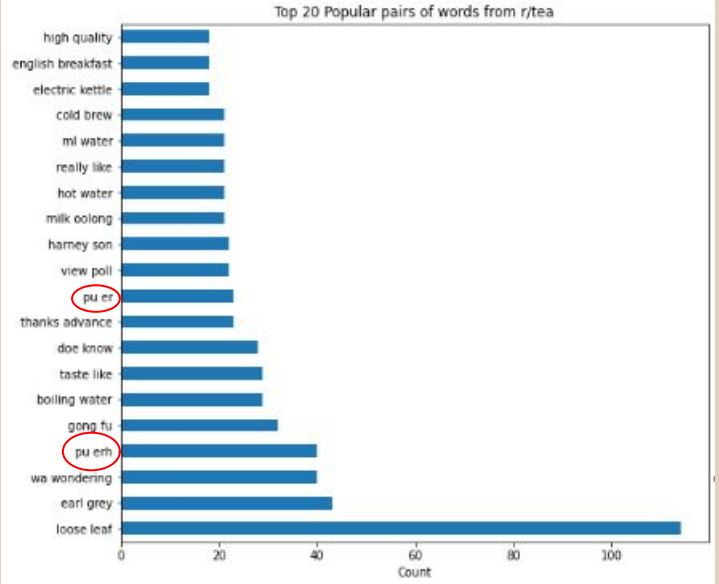
Further Improvements

Addressing Misspelling

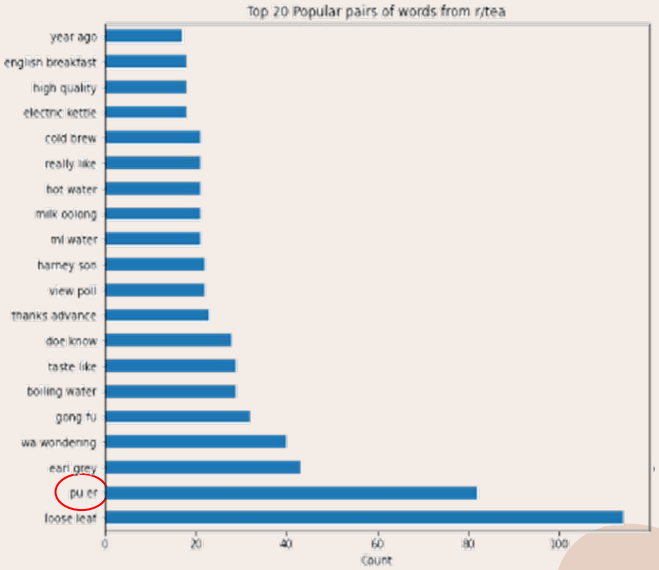
- Misspelled words not in the English dictionary are treated as separate occurrences of the word/ terms, e.g. pu’er vs pu’erh

- To create a dictionary to map observed misspellings

Before



After



Further Improvements

Addressing Multi-word names of tea, beans, and their equipments

- 1-gram does not pick up on some commonly observed, important terms that do not make sense as individual words such as earl grey

- Constantly update dictionary of singular words (removing spaces or hyphens between terms) so that future n-grams will yield more meaningful results

Further Improvements

Curating stopword dictionary

- Over iterations, dictionary of stopwords will become more comprehensive
- Model will better able to classify in subsequent runs

Further Improvements

Localized Source of Text Data

- Most reddit contributors are from the States and results may not localize well to the Asian context
- Constraint of project

Future Outlook

01

Other Social Media Platforms

Scrap and train data to be more robust

02

Other Languages

Tap into foreign markets

Future Outlook

03

Additional Offerings

With enough mentions, can offer other tea and coffee products

04

Engage in models catered for Visual Data

Especially for forum posts with low word counts

THANK YOU