Coffee, tea or .. ?

Subreddit Classification with NLP that classifies posts in r/Coffee and r/tea

- Marvin Fong

A little background...

Reddit is a social news website and forum where content is socially curated and promoted by site members through voting. The site name is a play on the words "I read it." and is composed of hundreds of subcommunities, known as subreddits. Each subreddit has a specific topic, such as technology, politics or music.

Introduction

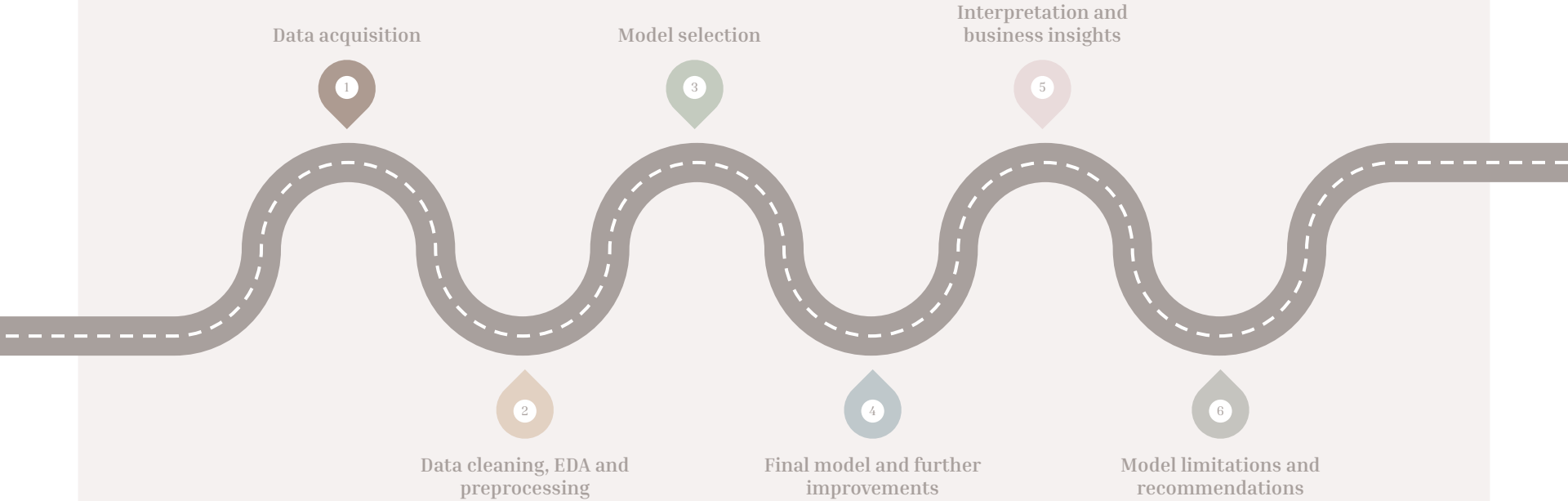
A marketing and advertising firm is planning to run a digital advertising campaign over the Christmas festive period. It aims to target tea and coffee lovers separately, despite having certain seemingly similar properties.

Problem Statement

Using a binary classification model to find some of the commonly used words among coffee and tea drinkers and recommend words that the marketing algorithm should pick up when people type these words on the internet, allowing targeted digital ads to show up on their social media feed,

This project aims to predict top predictive coffee and tea words from subreddit posts from r/Coffee and r/Tea respectively.

Agenda


Data acquisition
1
Data cleaning, EDA and preprocessing
2
Model selection
3
Final model and further improvements
4
Interpretation and business insights
5
Model limitations and recommendations
6

Data Acquisition

- Text data will be extracted from two different subreddits: r/tea and r/Coffee
- Utilised Pushshift Reddit API
- This API was created by the r/datasets moderator team to help provide enhanced functionality and search capabilities for searching Reddit comments and posts

Data Acquisition

The most recent 1,500 posts from each subreddit were extracted.

Example of posts scraped (from r/Coffee):

selftext	send_replies	spoiler	stickied	subreddit	subreddit_id	subreddit_subscribers	subreddit_type	thumbnail	title
Ive been enjoying coffee more in general and b...	True	False	False	Coffee	t5_2qhze	860476	public	self	new coffee enjoyer looking for suggestions one...
[https://youtu.be/UQV0J-lgcyE](https://youtu.b...	True	False	False	Coffee	t5_2qhze	860447	public	self	99% - 100% people Do Not Know... Tell Your Bud...
Ok so this may sound dumb, but whenever I make...	True	False	False	Coffee	t5_2qhze	860427	public	self	My coffee tastes like tomato?? 🍎☕
I was gifted an espresso machine for Christmas...	True	False	False	Coffee	t5_2qhze	860426	public	self	My coffee tastes sour. ☕

Data cleaning, EDA and Preprocessing

- Dropped moderator, deleted, duplicated, non-english removed posts, spam posts and promotional posts (only interested in actual posts)
- Stripped text of leading and trailing whitespaces
- Filled NaN values with blanks
- Removed URL links and HTML special entities

Data cleaning, EDA and Preprocessing

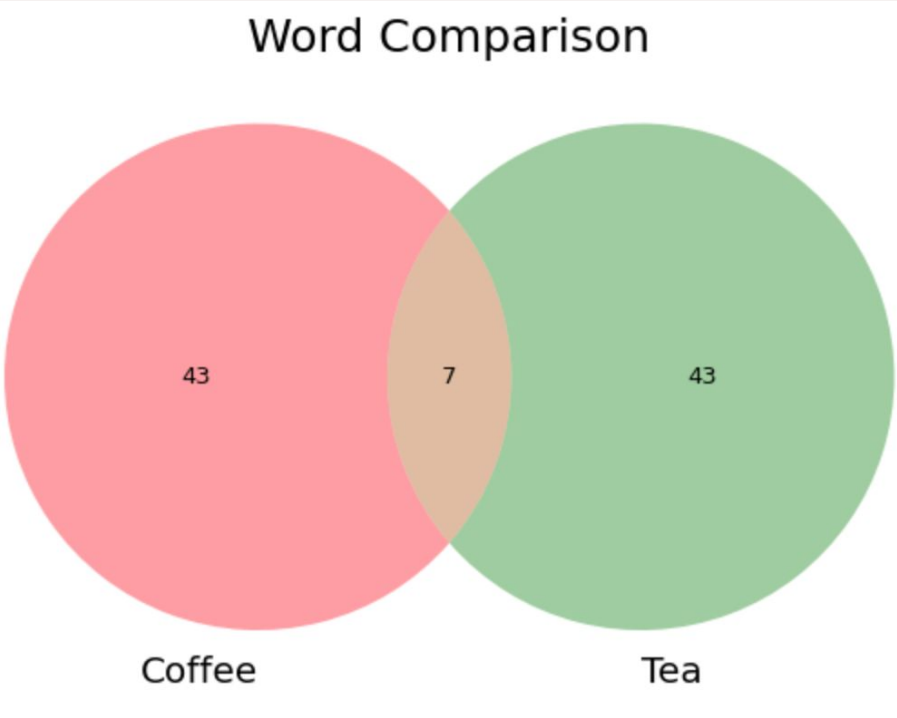
- Dropped all columns except ‘subreddit’, ‘selftext’ and ‘title’ columns
- Combined ‘selftext’ and ‘title’ into one column
- Under ‘subreddit’, binarized coffee and tea into ‘0’ and ‘1’ respectively

Most common words after preprocessing (Wordcloud)

r/Coffee

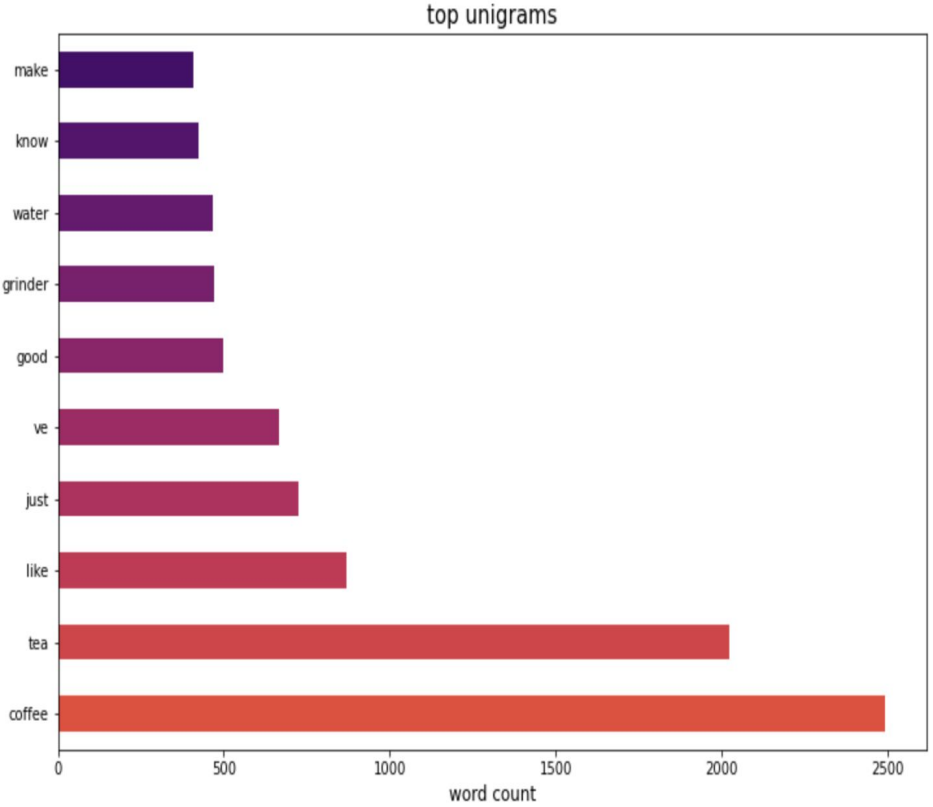
r/tea

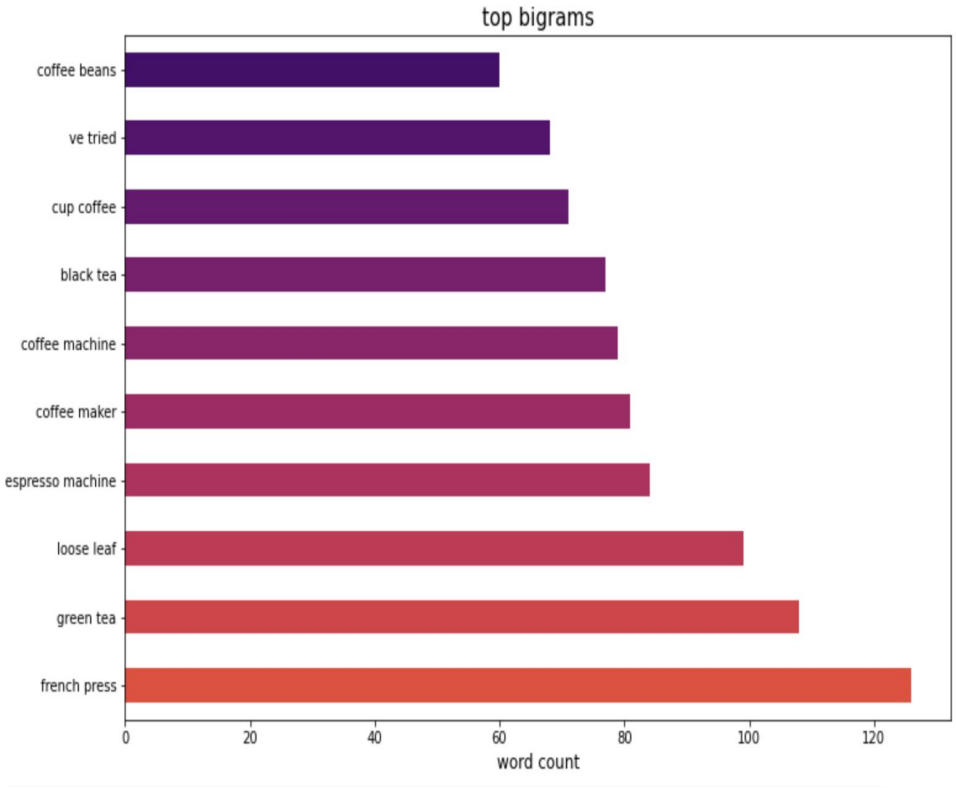
Most common words after preprocessing (Venn)



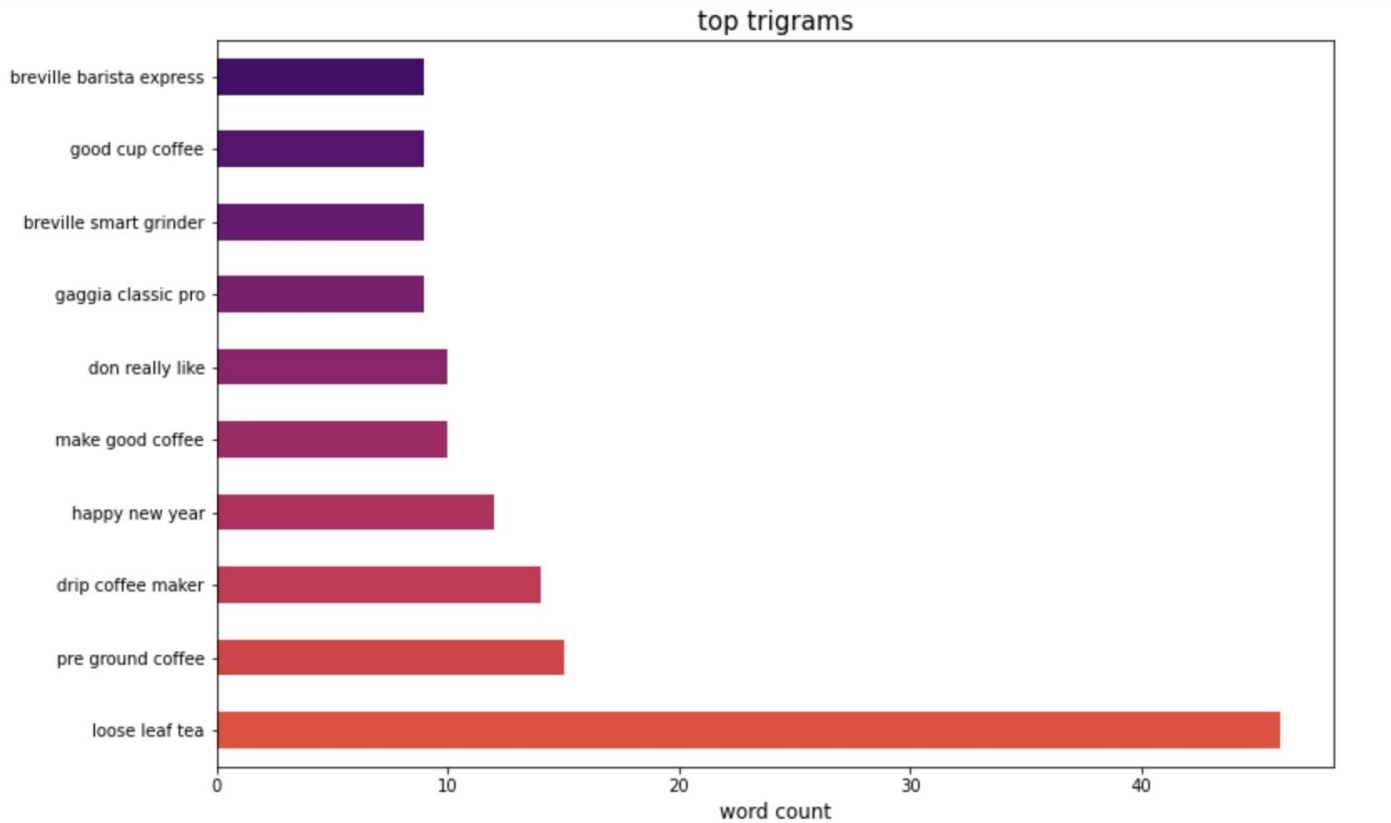
- Top 50 words for both subreddits
- 7 common words showing both subreddits to be different yet having certain similarities

Most common words after preprocessing (Ngrams)





Most common words after preprocessing (Ngrams)



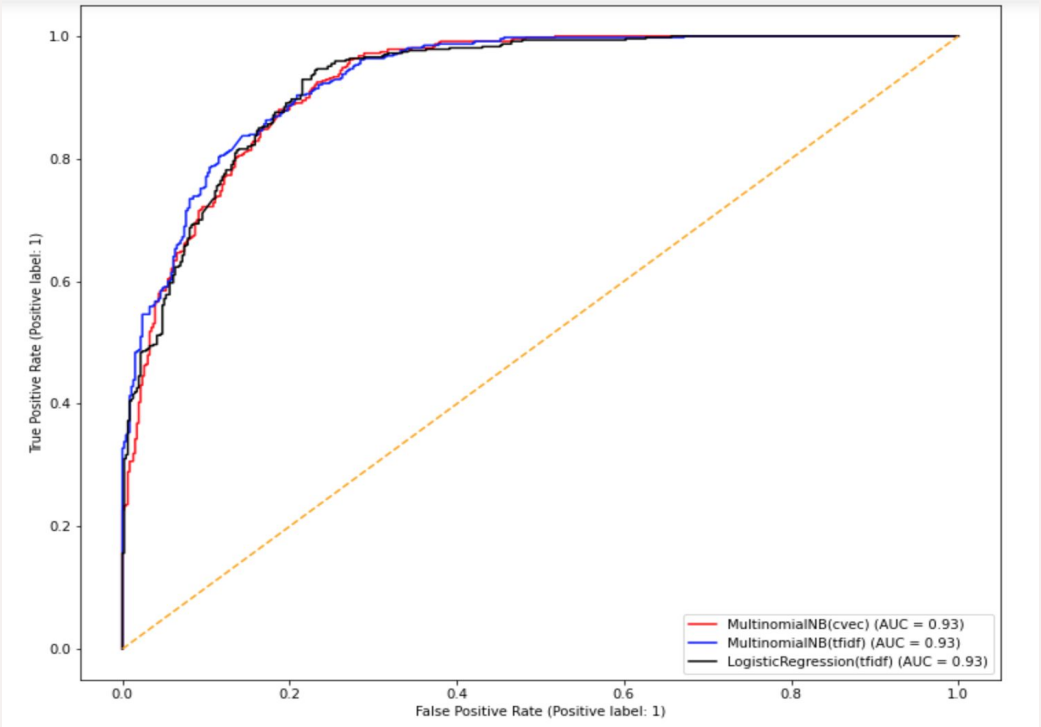
Model Selection

Various model performances:

	Model	Vectorizer	Train Score	Test Score	Best CV Score	False Positives	False Negatives
0	MultinomialNB	CountVectorizer	0.903243	0.842503	0.846363	86	60
1	MultinomialNB	TfidfVectorizer	0.961722	0.842503	0.850618	70	76
3	LogisticRegression	TfidfVectorizer	0.967039	0.842503	0.844234	98	48
6	ExtraTrees	CountVectorizer	0.987241	0.834951	0.819801	85	68
4	Random Forests	CountVectorizer	0.981925	0.833873	0.822448	105	49
7	ExtraTrees	TfidfVectorizer	0.989899	0.831715	0.820860	87	69
5	Random Forests	TfidfVectorizer	0.995215	0.830636	0.819777	107	50
9	SVM	TfidfVectorizer	0.992557	0.830636	0.838378	109	48
2	LogisticRegression	CountVectorizer	0.973418	0.827400	0.828294	115	45
8	SVM	CountVectorizer	0.869219	0.773463	0.754918	161	49

Model Selection

Various model performances:



Final model and further improvements

Final model chosen at this point:

- Multinomial Naïve Bayes model
- CountVectorizer (max_df = 0.5, min_df = 2, stop_words = custom stop words)
- Dataframe used: ‘subreddit’, ‘words’
  - ‘words’ = title + post content for both r/Coffee & r/tea
  - test_size = 0.33, stratify = y, random_state = 42

Result: Train Score = 90.3%

Test Score = 84.3%

False Predictions = 146 posts

Final model and further improvements

Original Model

Train Score	Test Score	False Positives	False Negatives
0.903243	0.842503	86	60

After Lemmatization (WordNetLemmatizer)

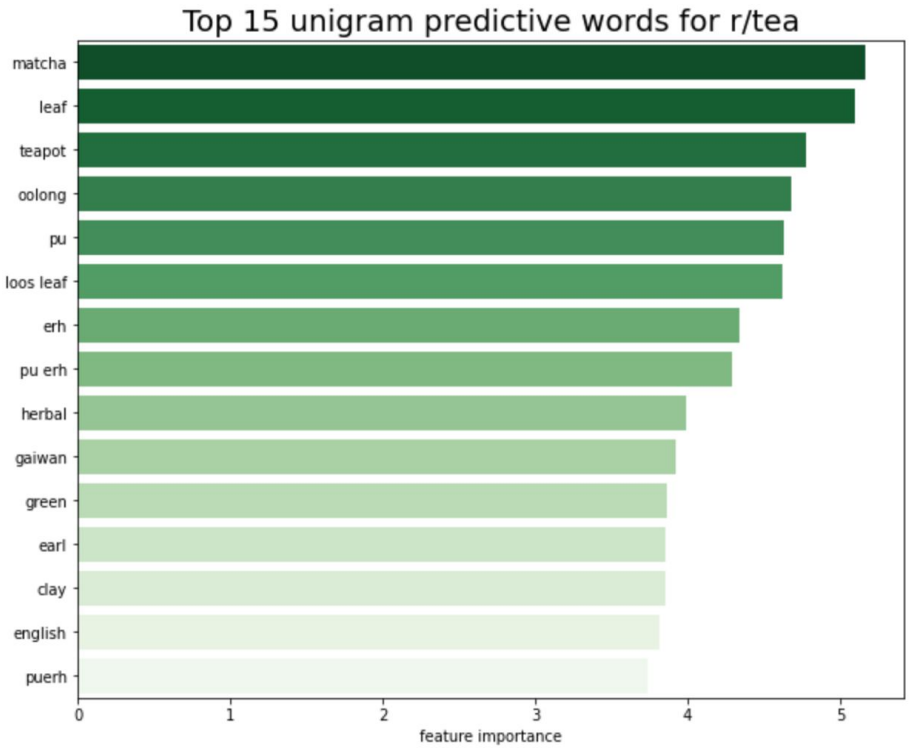
Train Score	Test Score	False Positives	False Negatives
0.905369	0.842503	87	59

After Stemming (Snowball Stemmer)

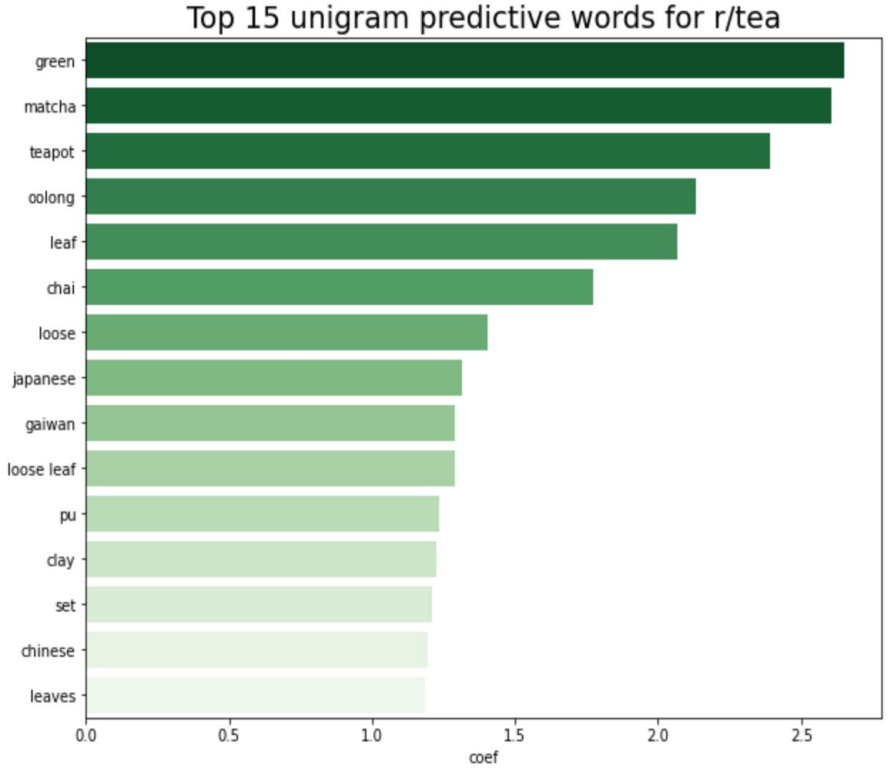
Train Score	Test Score	False Positives	False Negatives
0.919724	0.883495	53	55

Final model interpretation and business insights

Final Model

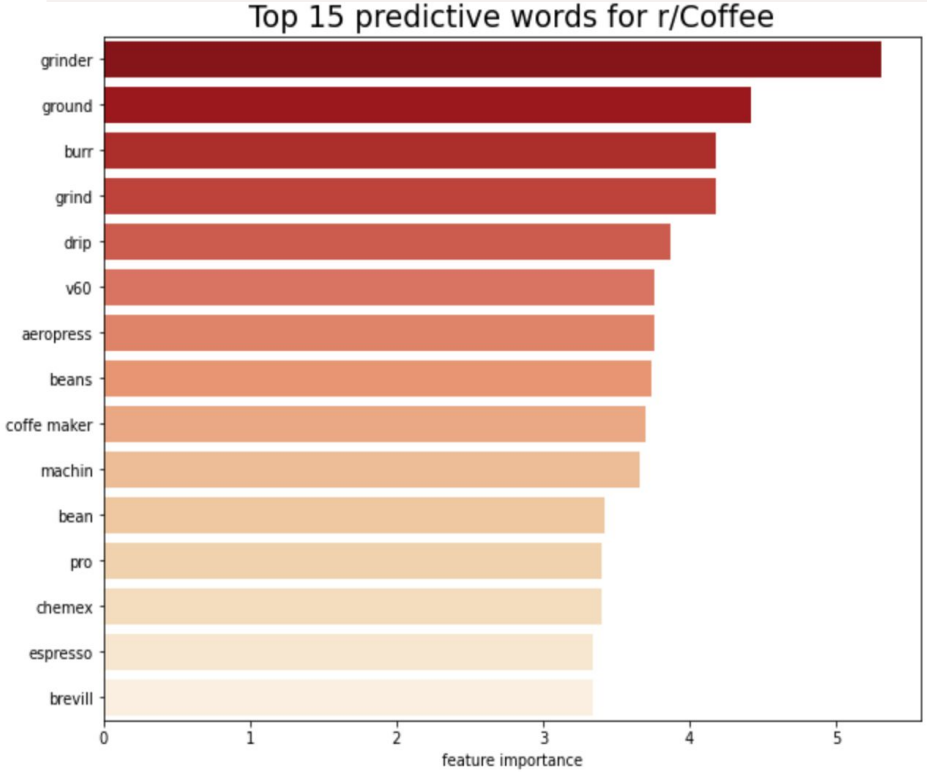


Logistic Regression Model

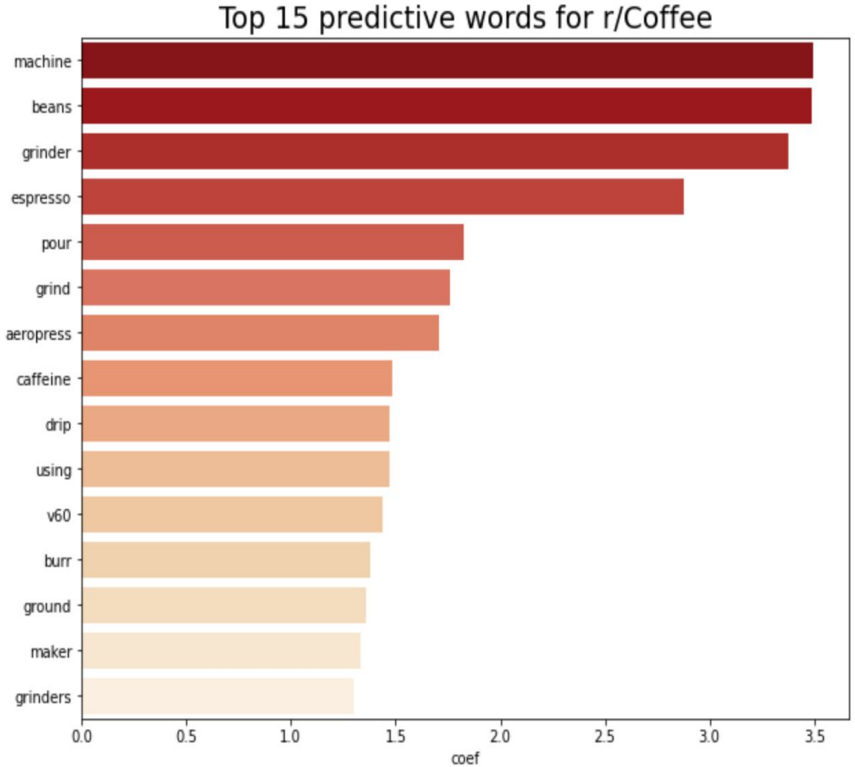


Final model interpretation and business insights

Final Model



Logistic Regression Model



Model limitations and recommendations

Limitations of model:

- Model predicted posts to be coffee instead of tea (False negatives)
- Sample of 5 were chosen for analysis:

```
made a fresh pot of tea

thought on whi peopl are less into light roast / oxid season tea.

can you use matcha in an espresso puk/machin ?

the right cup of tea!

breakfast at dunk coffe in sevilla,tea and donut
```

Model limitations and recommendations

Limitations of model:

- Model predicted posts to be tea instead of coffee (False positives)
- Sample of 5 were chosen for analysis:

dream hous

rate me. and be as harsh as you like

what is your feel toward vietnames milki coffee? i a/ i love it b/ it is just okay, mehhh c/ i heard about it but never tri befor d/ i don't even know if that is a thing e/ i dislik it

99% - 100% peopl do not know... tell your buddi lol [

how can i still enjoy coffe whilst be intolerant? a littl background: i absolut love coffee, most for the tast but also the cultur and energi boost. recent i have discov that i am probabl not intoler to caffeine, but to the bitter substanc within the coffee. at first i thought i was lactos intoler but my allergi test was negative, then i thought i might have a caffein intoler but i can consum everi type of caffein besid coffe and now i have last come to the conclus that i have to be intoler toward the bitter substances. right now i am feel sick, jitteri and fair exhausted, all becaus of one larg cup of cappuccino with milk. which realli sucks. my heart feel like it beat at 500 bpm (obvious exaggerated, but you get what i mean) and my stomach hurt and produc a lot of gases. it has been like this for the past few time when i enjoy my coffee. doe anyon have tip on how to still be abl to enjoy coffee? i realli don't want to give up on it. could chang the compani help? or the intensity? the dosis?? please, i'd liter do anyth (besid not drink coffe obviously). i'm sure someon here feel my pain and mayb know how to help me, so thank yall in advanc :)

Model limitations and recommendations

Recommendations for future models:

- a bigger corpus that incorporates a larger set of vocabulary on the topics of coffee and tea. This could also be taken from other sites such as food review blogs and related Facebook groups
- as mentioned earlier, preferences in the Food & Beverage scene are ever-changing. In order for our model to maintain a comparatively high accuracy, it should ideally be re-trained at regular periods so that it does not contain out-of-date information and trends from coffee/tea drinkers, for example the 'Dalgona coffee' craze that took place at the start of COVID-19

Model limitations and recommendations

Recommendations for future models:

- try other estimators like AdaBoost / GradientBoosting and try other vectorizers like lancaster Stemmer
- explore relationship between post content, number of comments, and upvote ratio
- use VADER (Valence Aware Dictionary and sEntiment Reasoner) lexicon to analyze the sentiments of posts

Thanks!

Now ... Coffee, Tea or any questions for Me?