Marketing and Advertising to Reddit Posters

How to Target Users Based on SubReddit Post Titles

Problem: Reddit offers advertising services on their platform. How do companies determine which subreddits to advertise their products? This study attempts to differentiate between subreddit accounts by the title of their posts to target users for advertisements.

Two Subreddits of recent interest were chosen for title analysis and prediction: Plant Based Diet and Keto

vs

Data Acquisition

PRAW (Python Reddit API Wrapper) → Pull down 1000 SubReddit posts from each SubReddit and from each extension: '.controversial', '.hot', and '.new' → Create dictionary from SubReddit posts and then a DataFrame →
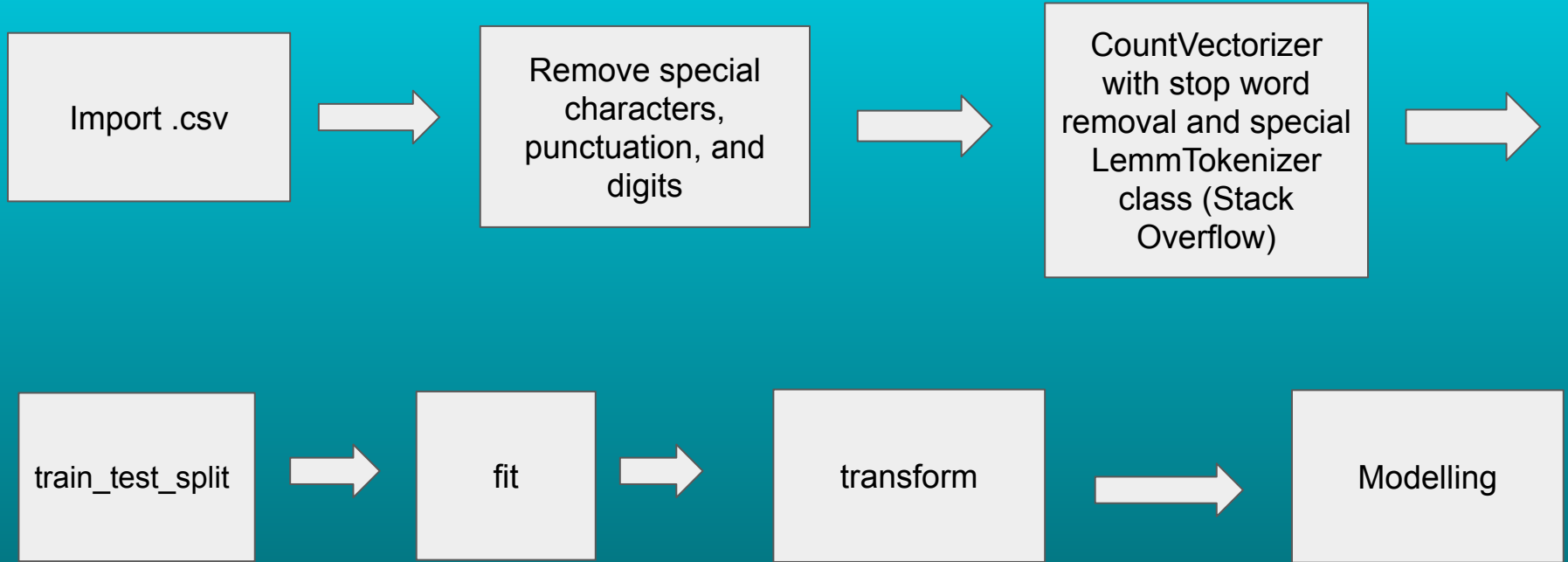
Create column designating 'Plant Based Diet' = 0 and 'keto' = 1 → Concatenate DataFrames/5653 posts/5159 after the duplicates were dropped → Create a .csv file

Data Cleaning

Import .csv → Remove special characters, punctuation, and digits → CountVectorizer with stop word removal and special LemmTokenizer class (Stack Overflow) →

train_test_split → fit → transform → Modelling

Single Word Analysis

word	count
keto	575
lb	288
diet	217
month	186
pic	184
plant	156
im	155
based	147
year	139
day	138
food	136
pound	134
just	124
weight	116
lost	107
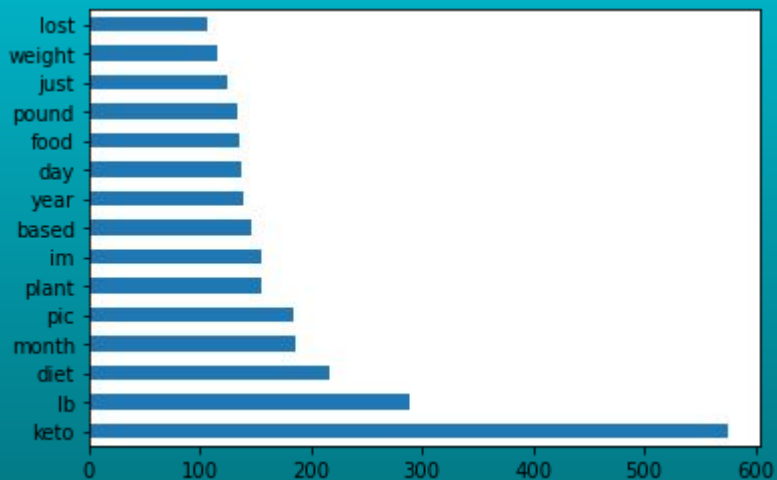
Bigram Analysis
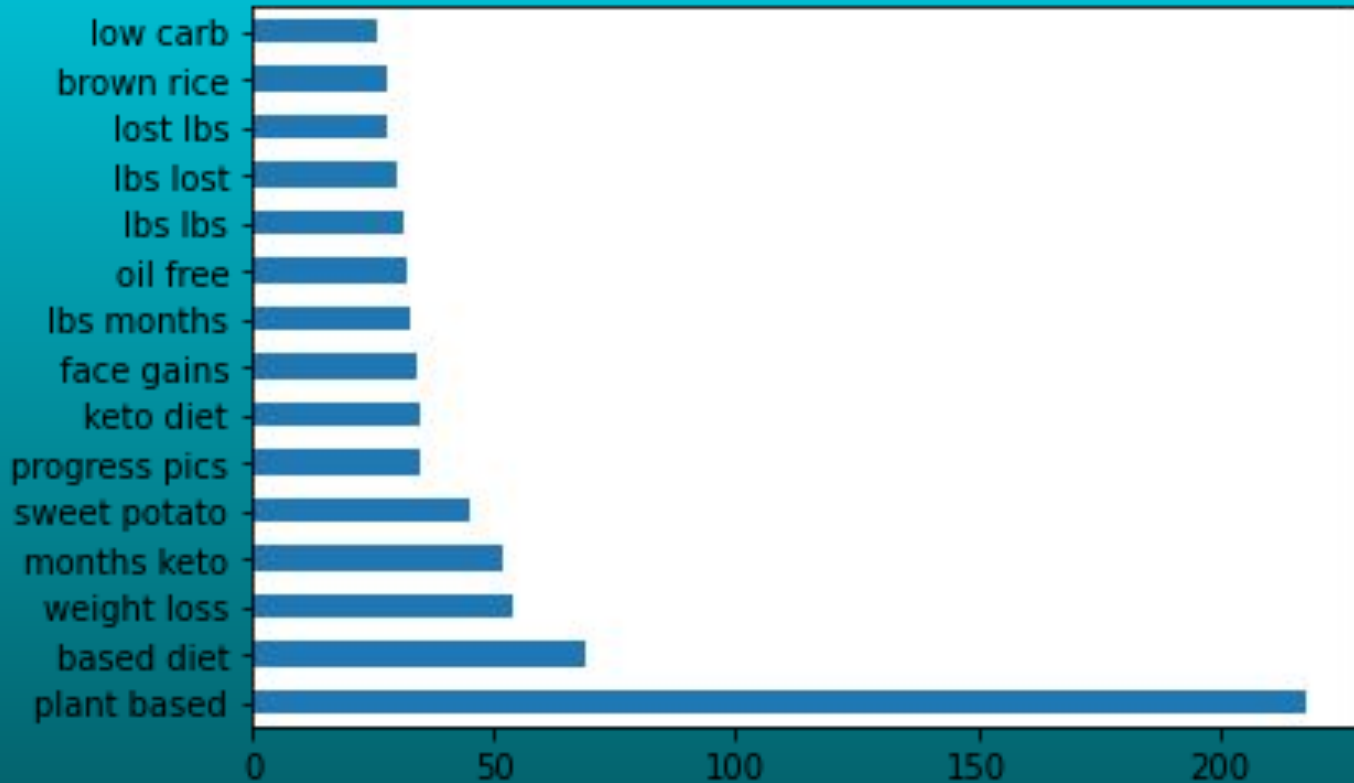
Trigram Analysis
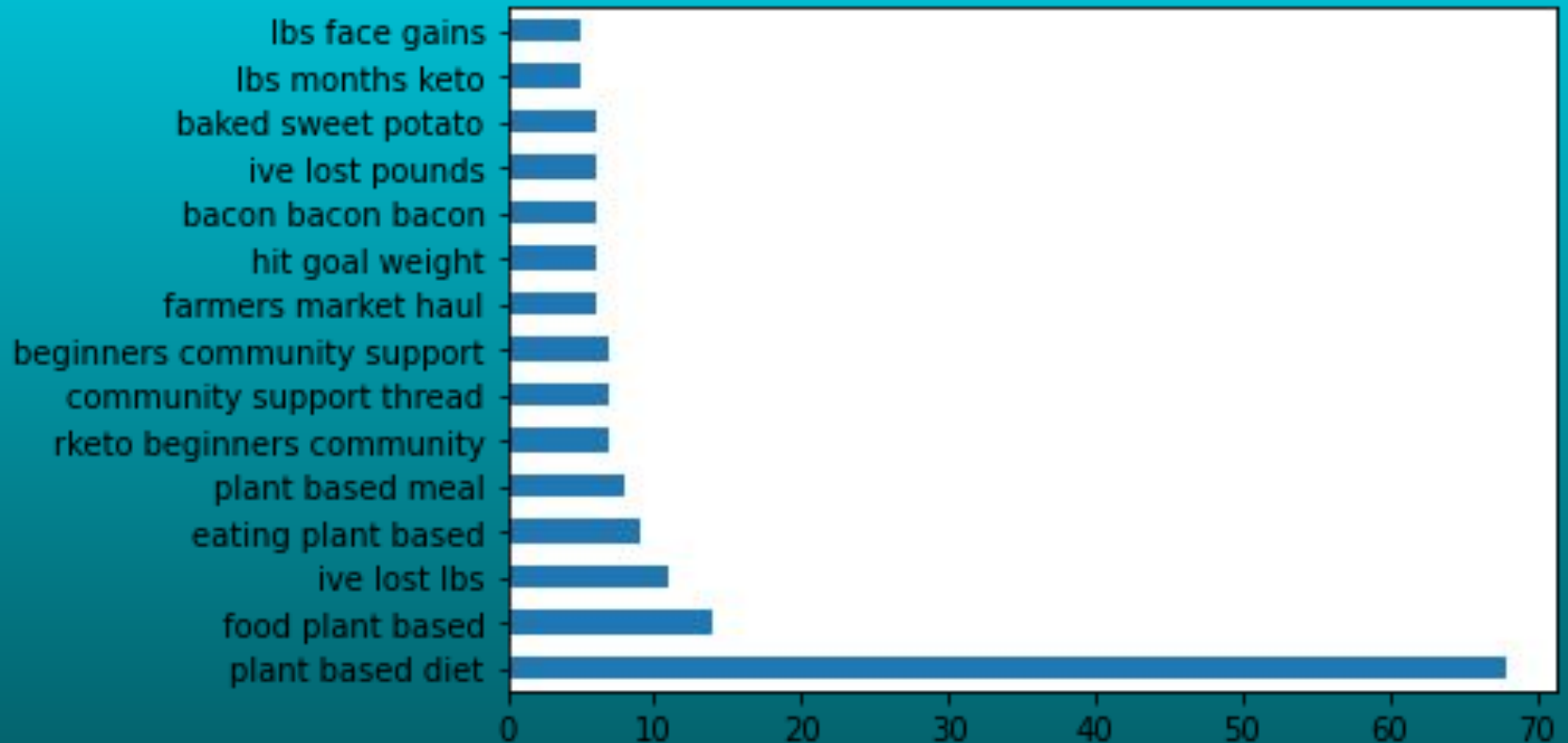
Models:

Naive Bayes | *Default Parameters*

Logistic Regression with regularization | *alpha = 0.01*

Random Trees Classifier | *max_depth = None, n_estimators = 200*

Extra Trees Classifier | *Default Parameters*

Classification Metrics

Plant Based Diet = 0 | keto = 1

Model	TN	FP	FN	TP
Naïve Bayes	652	175	117	759
Logistic Regression	632	195	127	749
Random Forest	634	193	124	752
Extra Trees	643	184	142	734

Classification Metrics

Plant Based Diet = 0   keto = 1

Model	Accuracy	Specificity	Sensitivity	Precision
Naïve Bayes	0.82	0.78	0.86	0.81
Logistic Regression	0.81	0.76	0.85	0.79
Random Forest	0.81	0.76	0.85	0.79
Extra Trees	0.81	0.77	0.83	0.79

Conclusions:

- All models performed similarly

- Accuracy for all models around 80 %

- For this study, specificity would have been preferred as false positives or plant based titles categorized as keto, are worse than the false negatives.

Further Study:

- Analysis of key words in posts would likely yield better accuracy and possibly less false positives.
- Further bigram and trigram analysis
- Sentiment analysis of other subreddits to uncover potential customers.