

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



<https://www.bing.com/images/search?view=detailV2&ccid=oQDRRD%2b&id=F5DEAA6AAB0061B1051E17EBD501AA566F54EB7F&thid=OIPoQDRD-u071FGW2uT-b1yghaE&mediaurl=https%3a%2f%2fplant-based.com%2fwp-content%2fuploads%2f2018%2f04%2fPlant-Based-Foods-EP9.jpg&expv=483&expw=724&q=plant+based+diet+images&simid=608010448563406751&ck=846421841C083FEC0A7E78793163AE1&selectedIndex=21&qv=plant+based+diet+images&FORM=IRPRST&ajaxhist=0>

VS

KETO

How people think I eat



How I really eat



<https://www.bing.com/images/search?view=detailV2&ccid=p6NfS5&id=94B4A8E7553437F395F128DC50ED50AB2C78E4D&thid=OIP.p6NfS5WuSdh1o7mlghw1aF&mediaurl=https%3a%2f%2frobwolf.com%2fwp-content%2fuploads%2f2017%2f12%2fketto-eating.jpg&expv=750&expw=1000&q=what+people+on+think+you+eat+on+keto+and+actually+eat+on+keto+image&simid=608032245464367209&ck=B3A1CEA1BC95542E318A6187E23A9E1B&selectedIndex=0&qv=what+people+on+think+you+eat+on+keto+and+actually+eat+on+keto+image&FORM=IRPRST&ajaxhist=0>

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

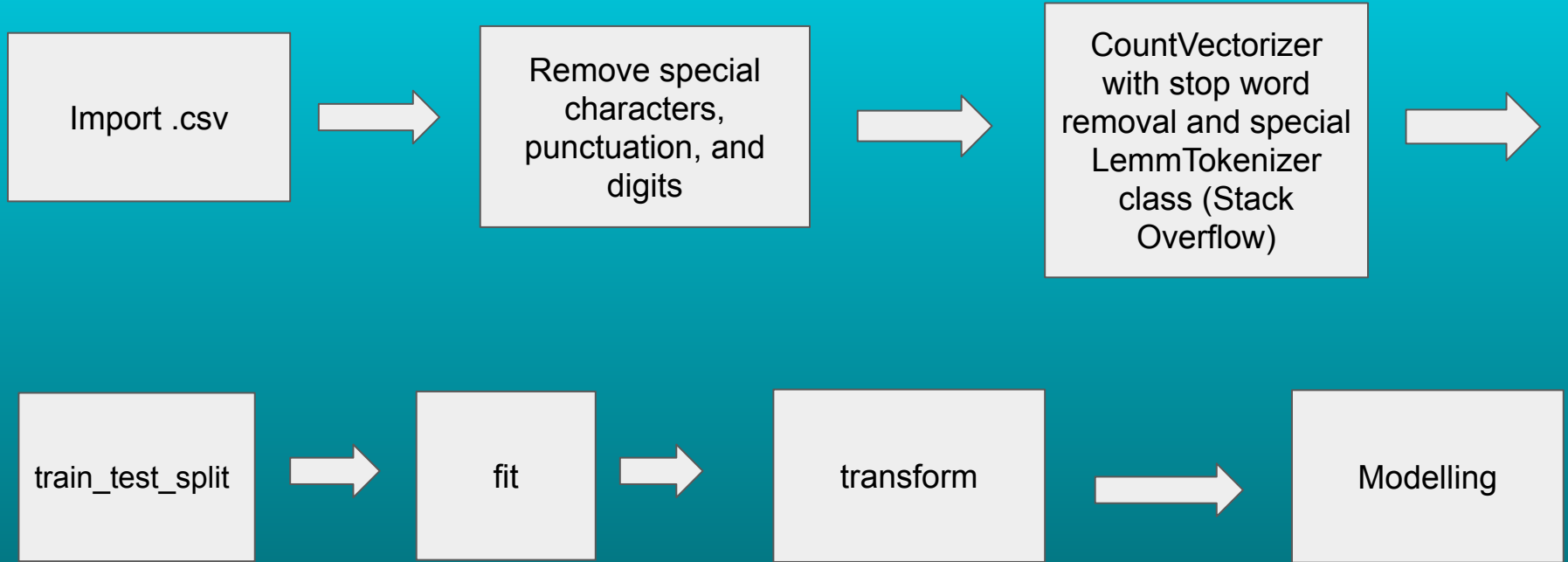


Concatenate
DataFrames/5653
posts/5159 after the
duplicates were
dropped



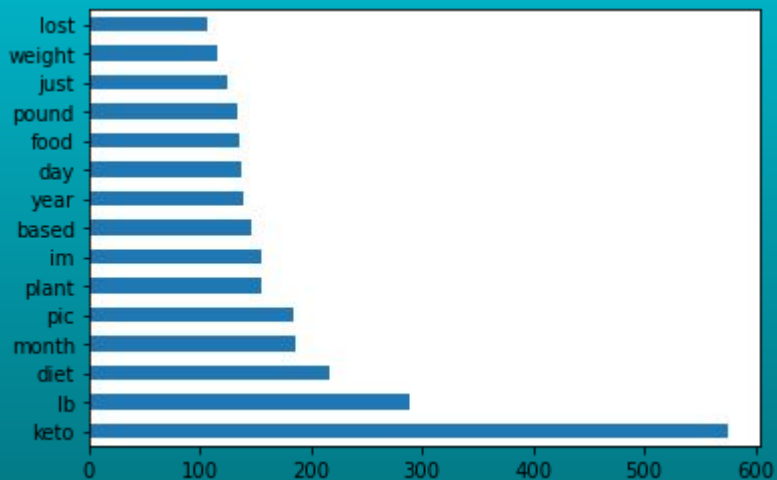
Create a .csv file

Data Cleaning

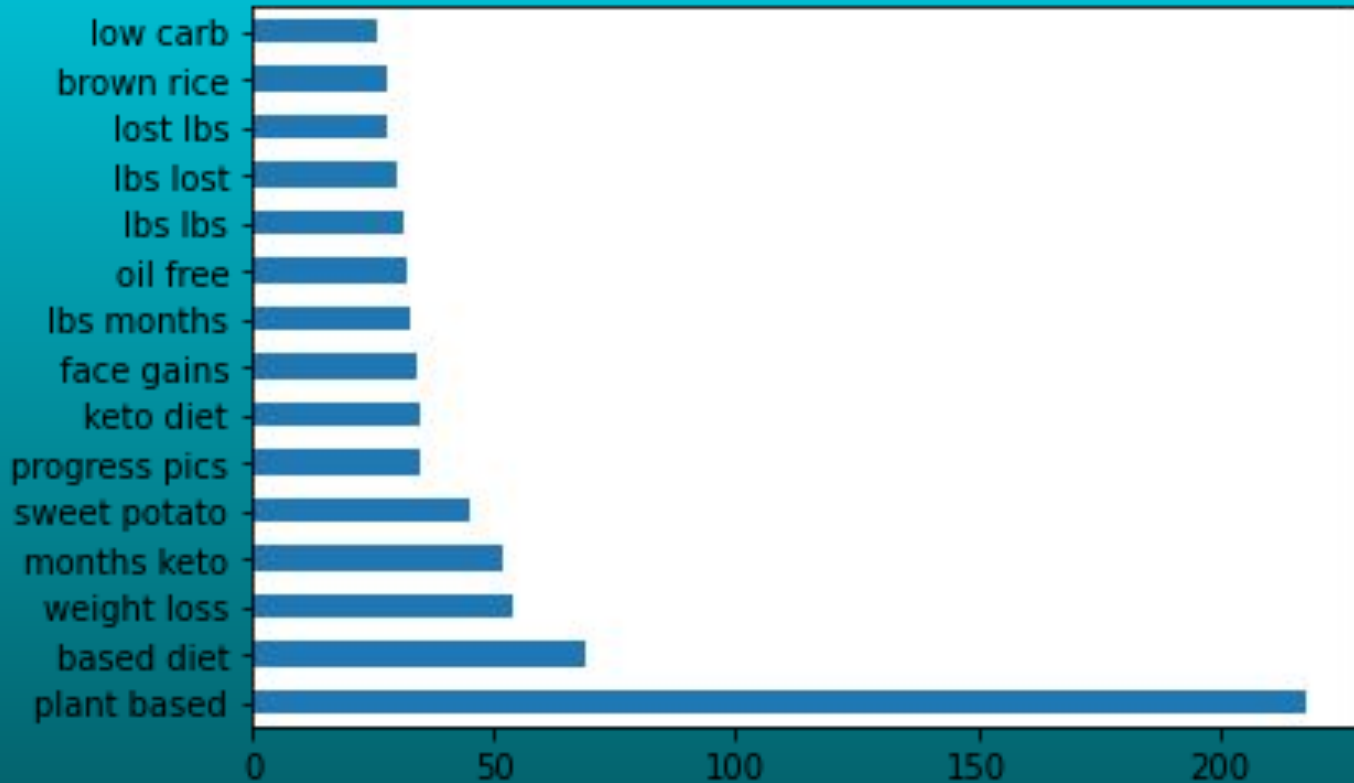


Single Word Analysis

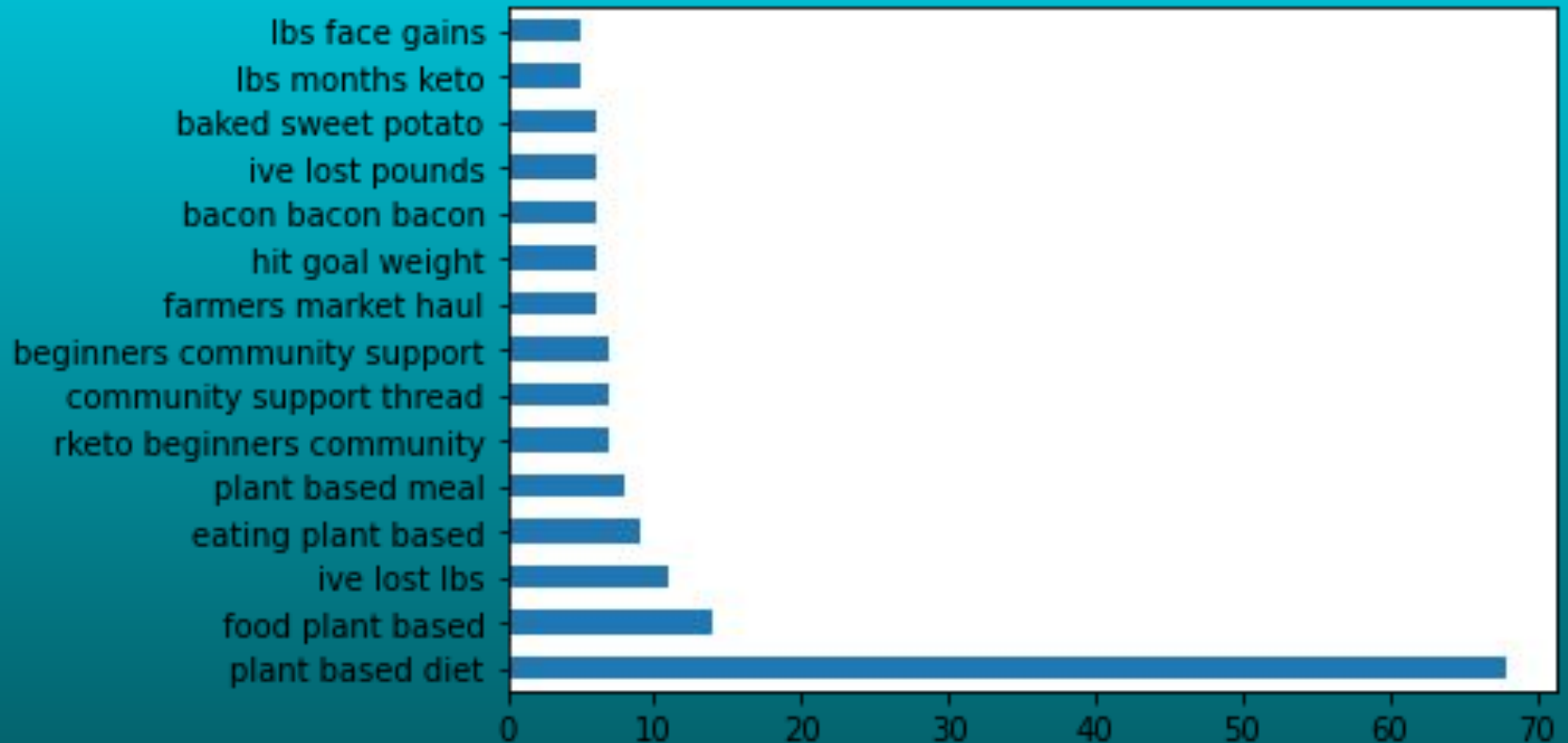
<i>word</i>	<i>count</i>
<i>keto</i>	575
<i>lb</i>	288
<i>diet</i>	217
<i>month</i>	186
<i>pic</i>	184
<i>plant</i>	156
<i>im</i>	155
<i>based</i>	147
<i>year</i>	139
<i>day</i>	138
<i>food</i>	136
<i>pound</i>	134
<i>just</i>	124
<i>weight</i>	116
<i>lost</i>	107



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

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

Classification Metrics

Plant Based Diet = 0 keto = 1

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

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.