Classification Model

r/vegan vs r/keto

Lloyd
Jean
Lucus

Table of Contents

Problem Statement

- 'Healthy Eats', a company whose mission is to create awareness on healthy diets for consumers, owns a forum with multiple diet categories.

- To create and train a model that can identify keywords and classify posts into the correct diet category taking reference from two popular subreddit, Vegan and Keto.

Methodology

Count Vectorizer
TF-IDF Vectorizer

Random Forest
Logistic Regression

Background

r/vegan

Animal Consumption Free Lifestyle

Goal: Abstain from participation in any form of animal cruelty

r/keto

Low Carbohydrate, High Fat Diet

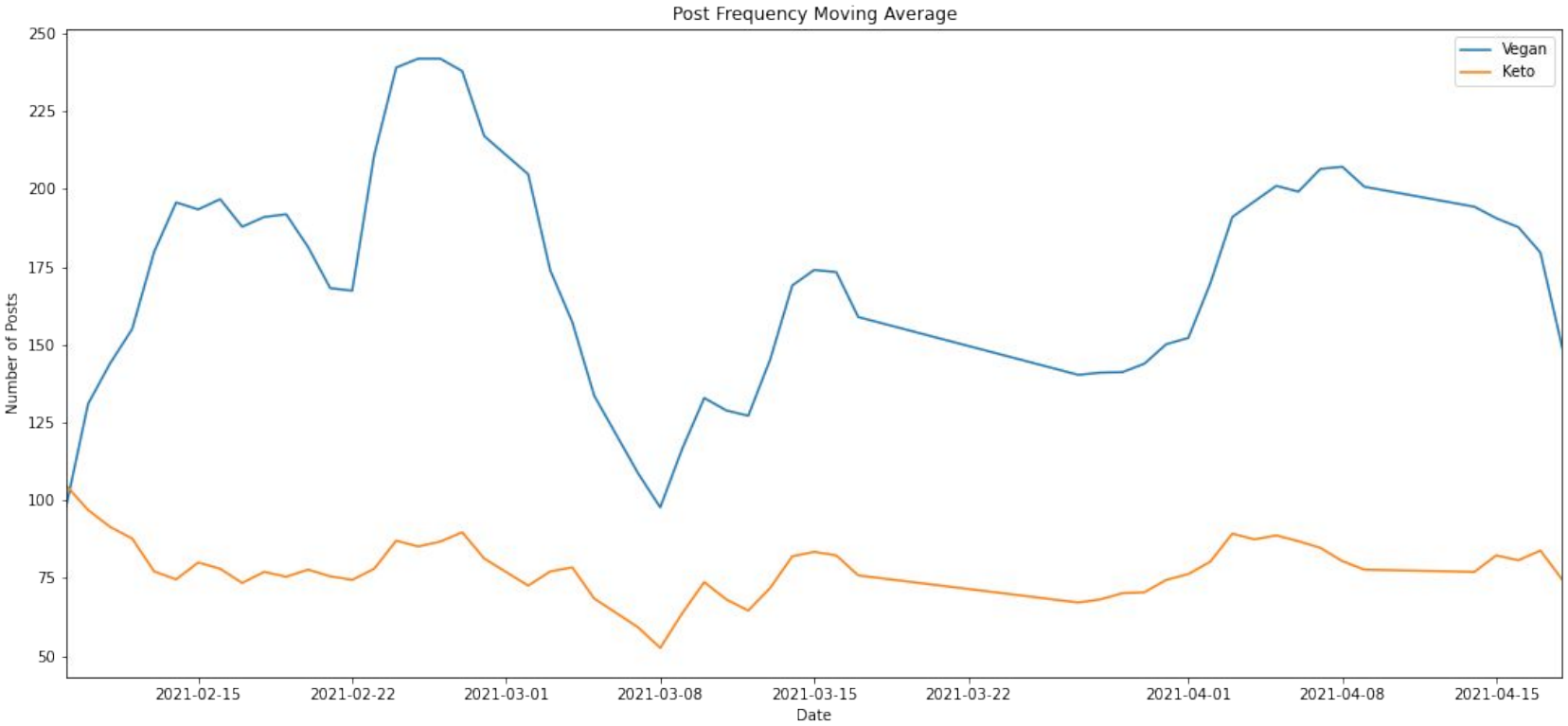
Goal: Lose weight fast

Table of Contents

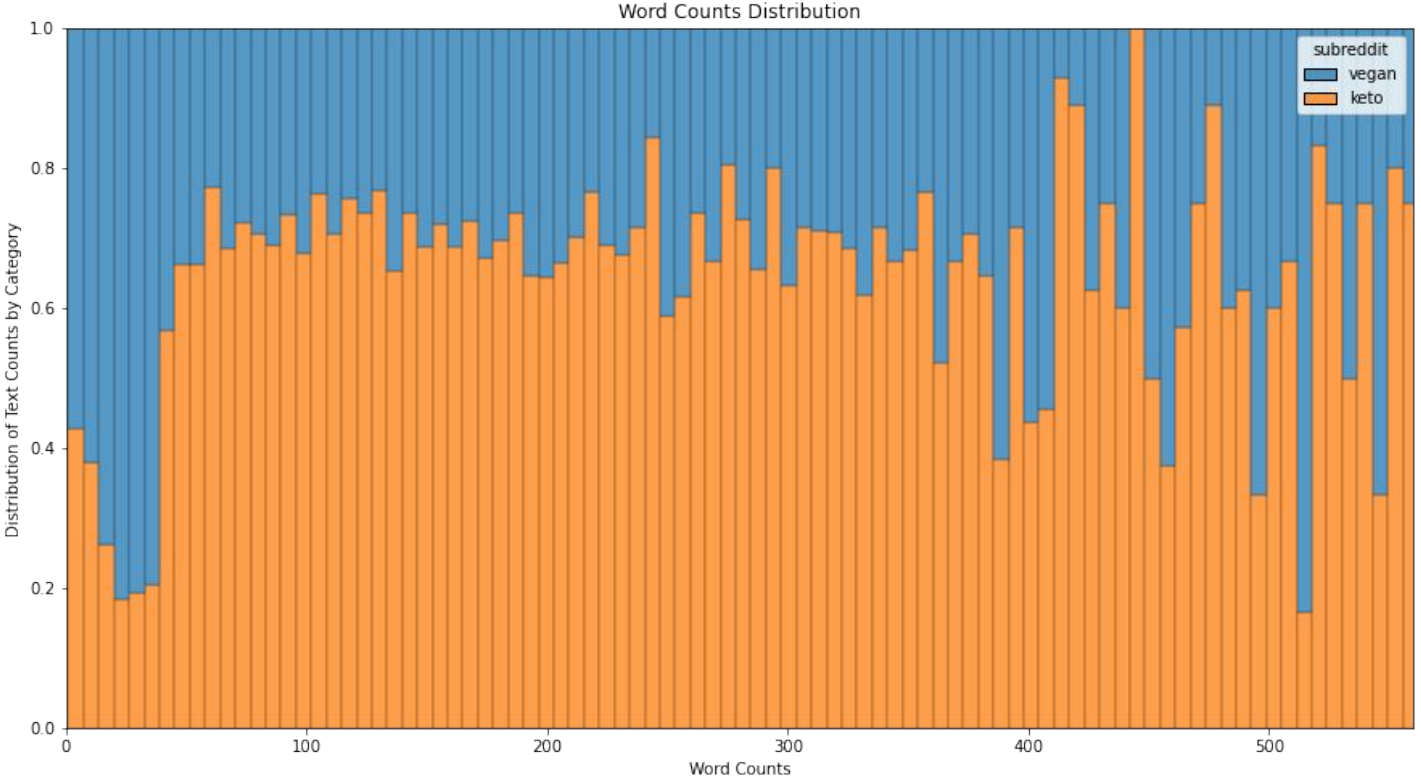
Exploratory Data Analysis - Metadata

	Created	No of Subscribers	AVG No of Posts per Day
r/vegan	31 Mar 2008	592,273	40.0
r/keto	27 May 2010	2,371,156	41.67

Exploratory Data Analysis - Posts Volume

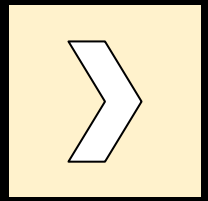


Exploratory Data Analysis - Word Counts



Exploratory Data Analysis - Flair Text / Tags

r/vegan

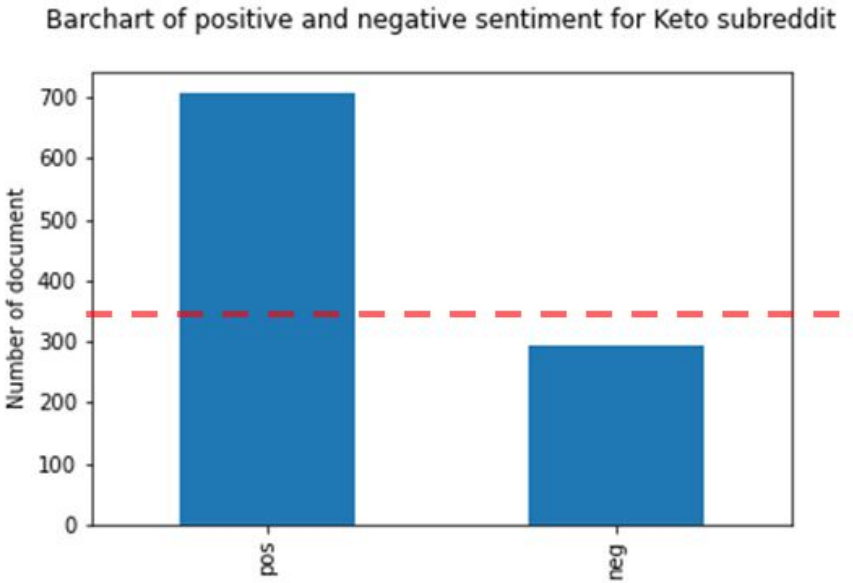


Food
Question
Discussion
Health

r/keto

Help
Food & Recipes
Success Story
Tips and Tricks

Exploratory Data Analysis - Sentiment Analysis



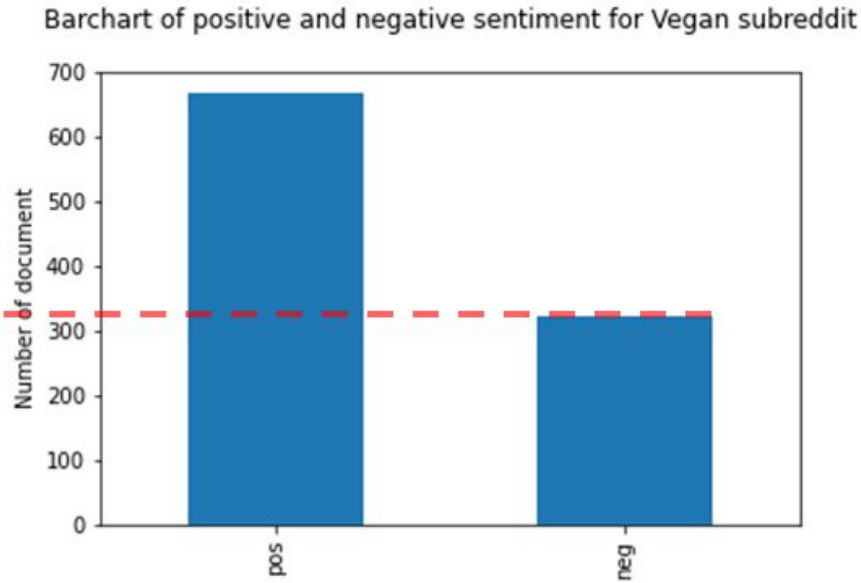


Table of Contents

Baseline Model

- Serves as the benchmark for our model to beat.
- The baseline accuracy is the percentage of the majority class.
- Baseline Accuracy: 50.9%

Transformers: Countvectorizer

- Enables the pre-processing of text data.
- Converts a collection of text documents to a vector of term/token counts. (1's & 0's)
- Builds a vocabulary of known words.

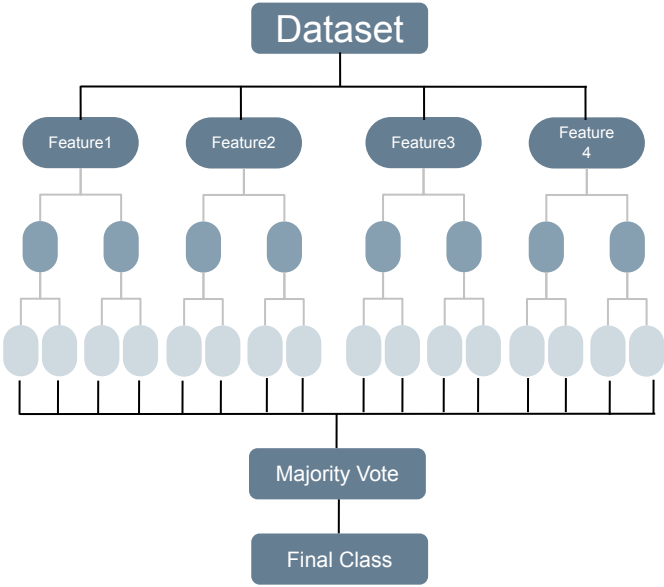
	the	red	dog	cat	eats	food
dog eats cat ⇨	0	0	1	1	1	0
the cat eats ⇨	1	0	0	1	1	0
red dog eats ⇨	0	1	1	0	1	0

Transformers: TF-IDF Vectorizer

- Statistic that aims to better define how important a word is for a document.
- A score which is applied to every word in every document in our dataset.
- TF-IDF value increases for every appearance of a word in the post
- And decreases with every appearance in other posts.

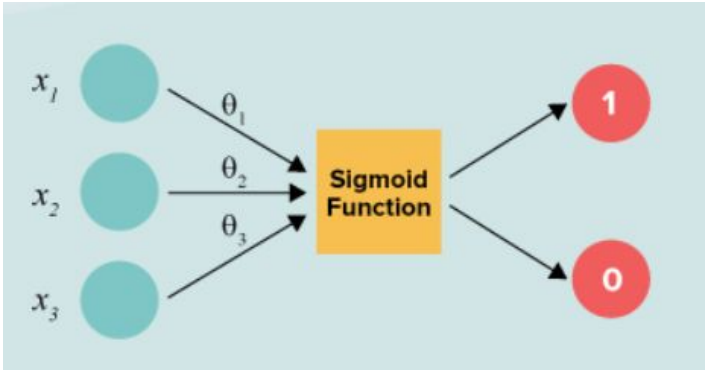
Model: Random Forest

- Classification Algorithm consisting of many decision trees.
- Uses bagging and feature randomness to build each tree
- Creating an uncorrelated forest of trees



Model: Logistic Regression

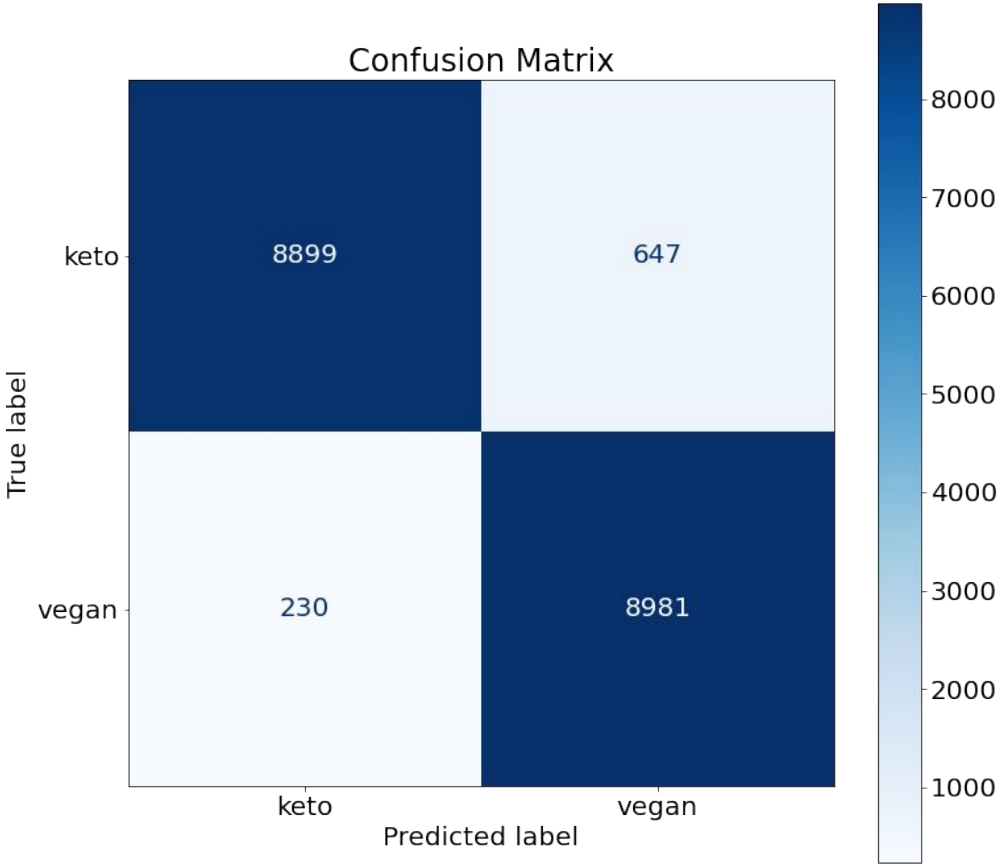
- Statistical model that uses Logistic function to model the conditional probability
- Variables can be numeric or categorical



Evaluation of Training Model

Models	Vectorizer	Accuracy	Train Score	Test Score
Baseline Model		0.509	-	-
Random Forest	Count Vectorizer	0.9358	0.9638	0.9358
Random Forest	TF-IDF Vectorizer	0.9398	0.9645	0.9398
Logistic Regression	Count Vectorizer	0.9366	0.9718	0.9366
Logistic Regression	**TF-IDF Vectorizer**	**0.9500**	**0.9590**	**0.9400**

Final Model - Logistic Regression with TF-IDF Vectorizer



Accuracy: 0.95

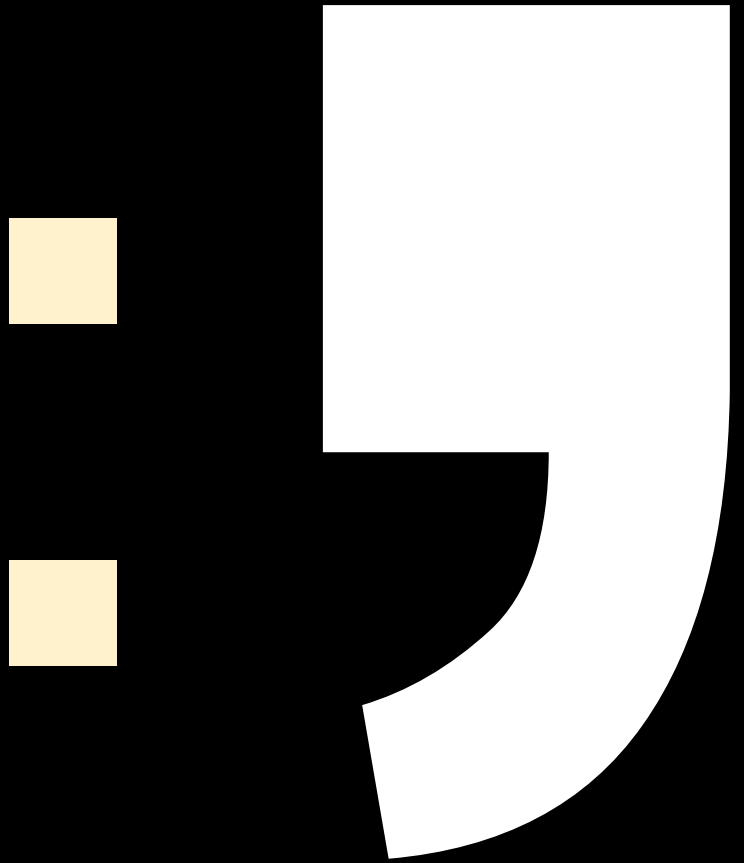
Sensitivity/Recall: 0.98

Specificity: 0.93

Precision: 0.93

Why is this important?

Type I Error: Keto posts are wrongly classified as Vegan aka False Positive



Top Keywords

r/vegan	r/keto
vegan	keto
animal	carb
meat	fat
plant	start
tofu	weight
seitan	sugar
plant based	macro
go vegan	fast

Table of Contents

Conclusions

- Based on observed best modelling scores:
    - Logistic Regression (TF-IDF Vectorizer) at **94%** test score
- Top keywords suggest a distinct separation between the two subreddits
- Other diets can be explored with the working model for further analysis

Recommendations

- Review on the misclassified posts keywords prior to postings in vegan forum
- Increase the time horizon of the data collected to capture lingos of different subreddits.

Thanks

Do you have any questions?

addyouremail@freepik.com
+91 620 421 838
yourcompany.com