Nutrition vs Keto

Subreddit Classification



Data Science Process

01 BUSINESS Problem and Introduction

02 DATA CLEANING and Pre-processing

03 EXPLORATORY Data Analysis

04 MODEL SELECTION and Insights

Business Problem:

Can we automate the classification of post so that only appropriate topics are posted for each subreddits?

Intro to our subreddits

Nutrition

A subreddit for the discussion of nutrition science. Macronutrients, micronutrients, vitamins, diets, and nutrition news are among the many topics discussed.

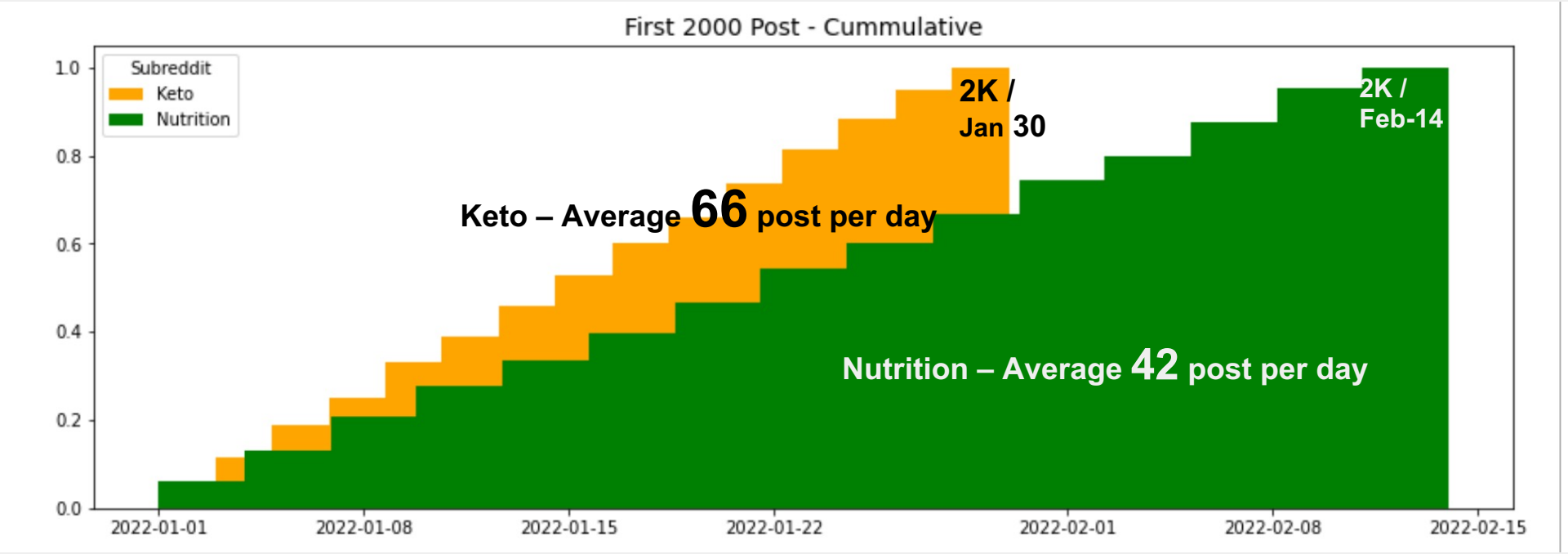
2.3M Members

Keto

The Ketogenic Diet is a low carbohydrate method of eating. Place to share thoughts, ideas, benefits, and experiences around eating within a Ketogenic lifestyle. Helping people with diabetes, epilepsy, autoimmune disorders, acid reflux, inflammation, hormonal imbalances, and a number of other issues, every day.

2.9M Members

Data Collection (Data from Jan to Feb '22)



First 2000 post from Jan 1, 2022:

Since Keto is more active and have more members, it reached the 2K post on Jan 30, 2022. Nutrition's reached 2K

Cleanup and Pre-processing

Cleanup Author:

Keto- >AutoModerator, Basic_Site5449 **1.5%**

Nutrition >AutoModerator Disastrous-Drop-8085 **2.0%**

Remove Duplicate:

Title **2.0%**

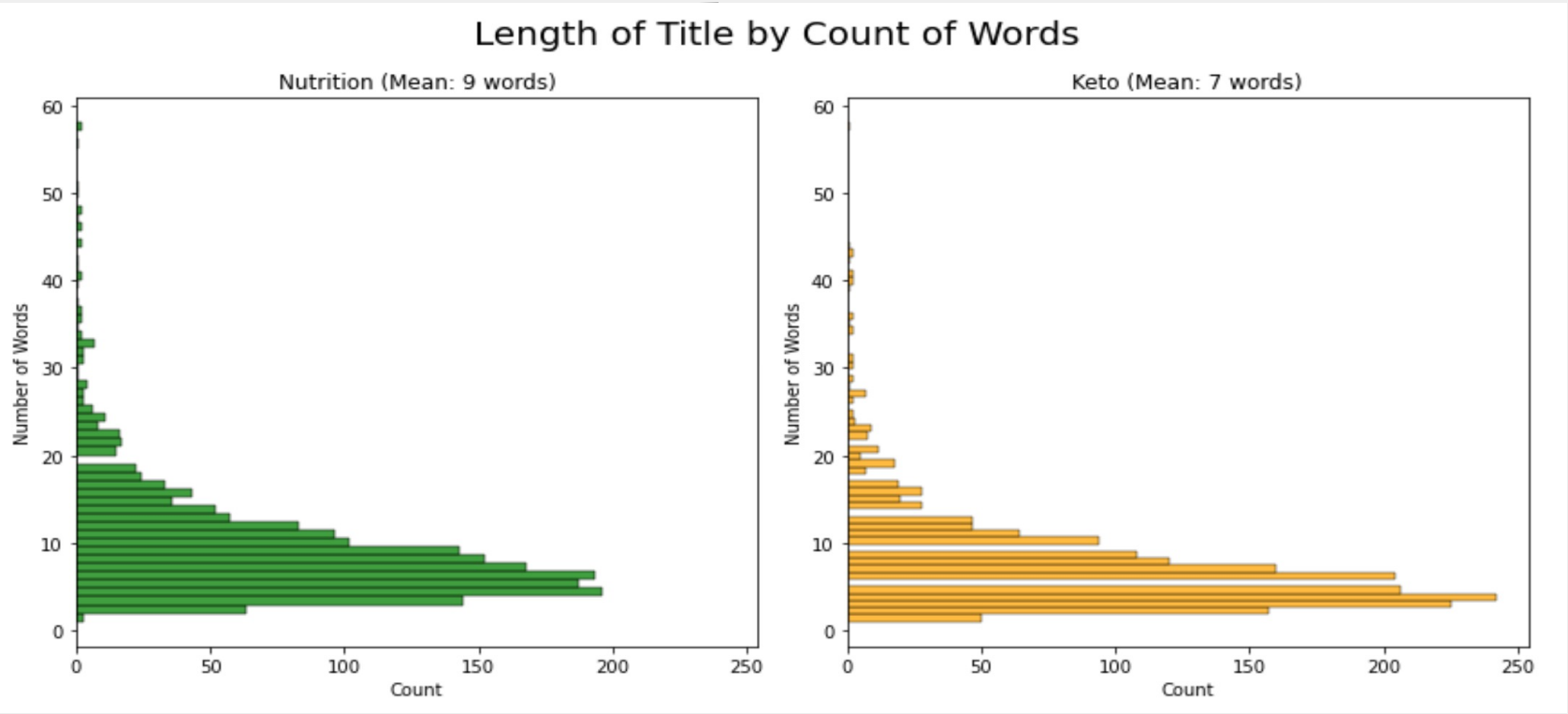
- **Tokenizing** (lower case and removed punctuations)
- **Stemming** (transforms a word into its root form, *e.g. eating/eats to eat, losing/loses to lose, why to whi, does to doe*)



Feature(s)?

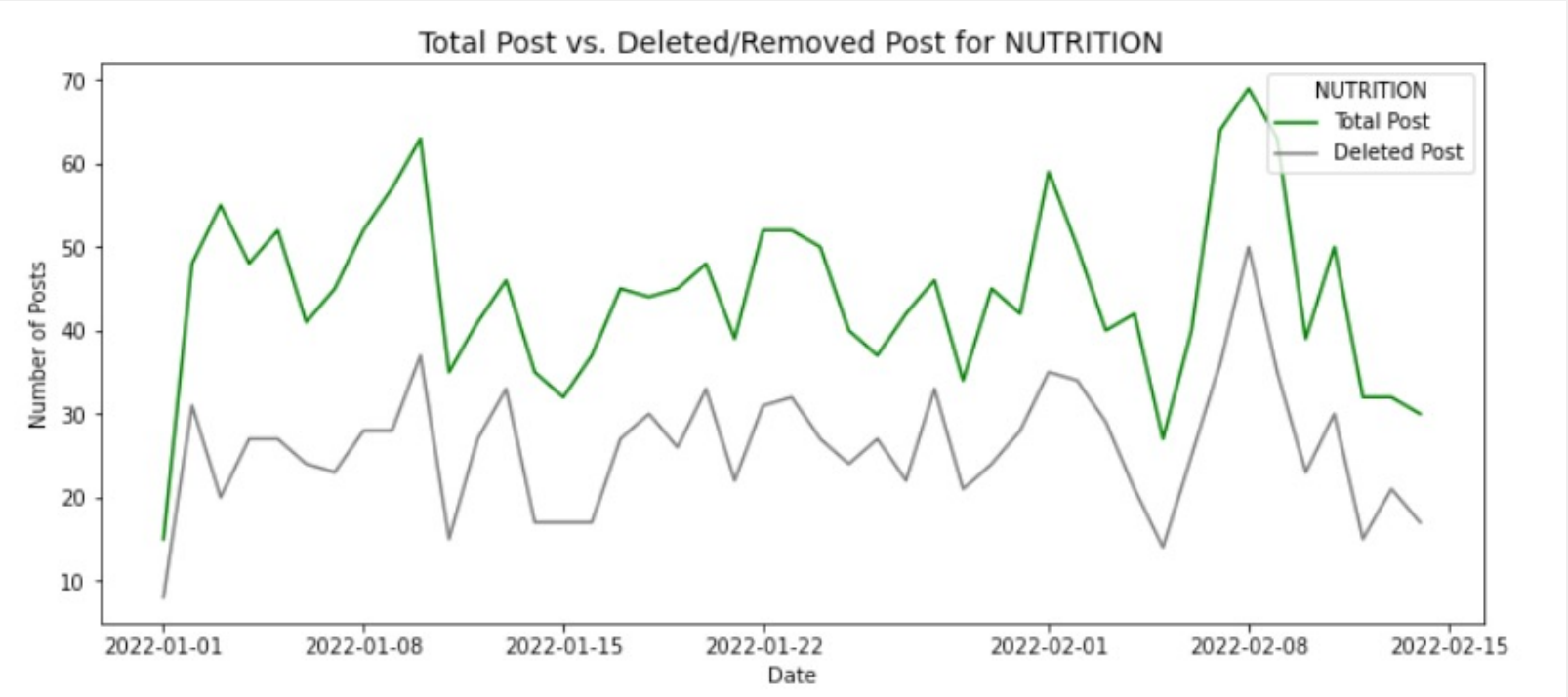
Title or Self-text,
... Or both?

Exploration/ Title Analysis



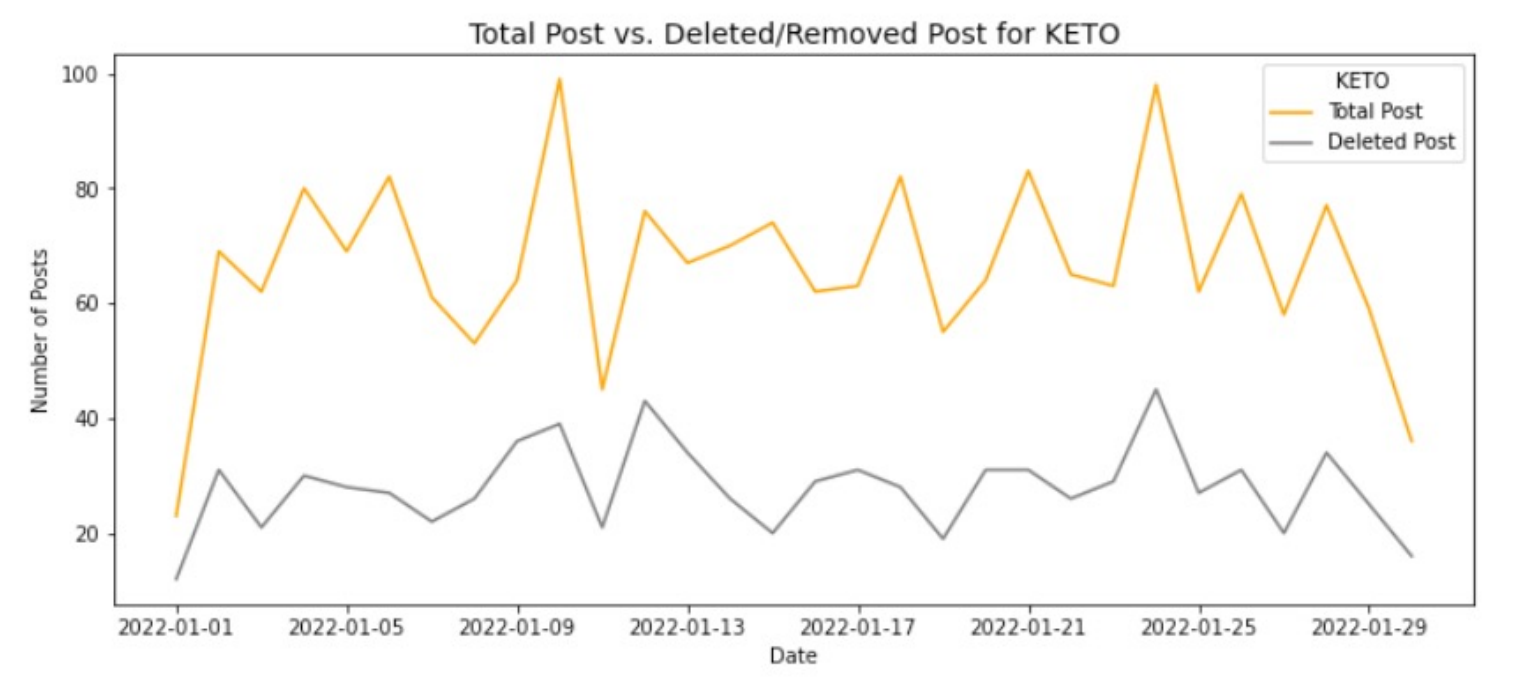
Nutrition has slightly higher average count of words, they have (almost) similar distributions

Self-Text Analysis



59% of the self-text was removed/deleted

Self-Text Analysis

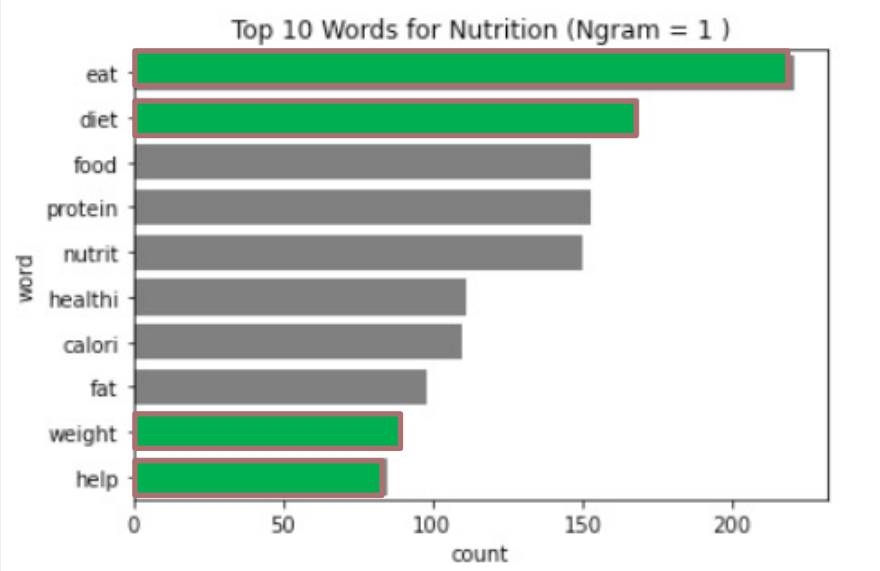


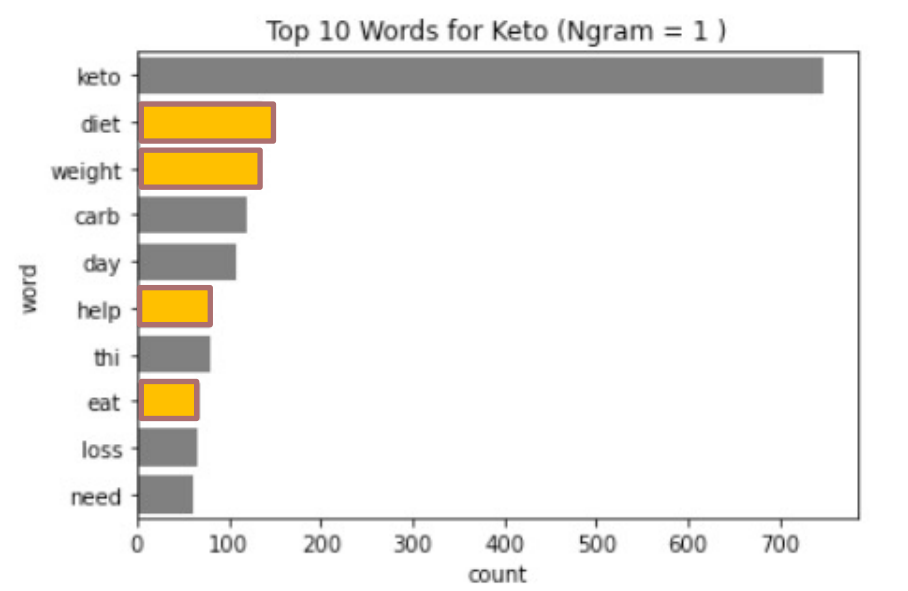
41% of the self-text was removed/deleted

Feature: Title

We use **CountVectorizer** to extract feature from ***Title..*** to transform ***Title*** into a bag of words (in a simple term), or a vector on the basis of the frequency or count of each word that occurs in the entire ***Title*** (as a technical description)

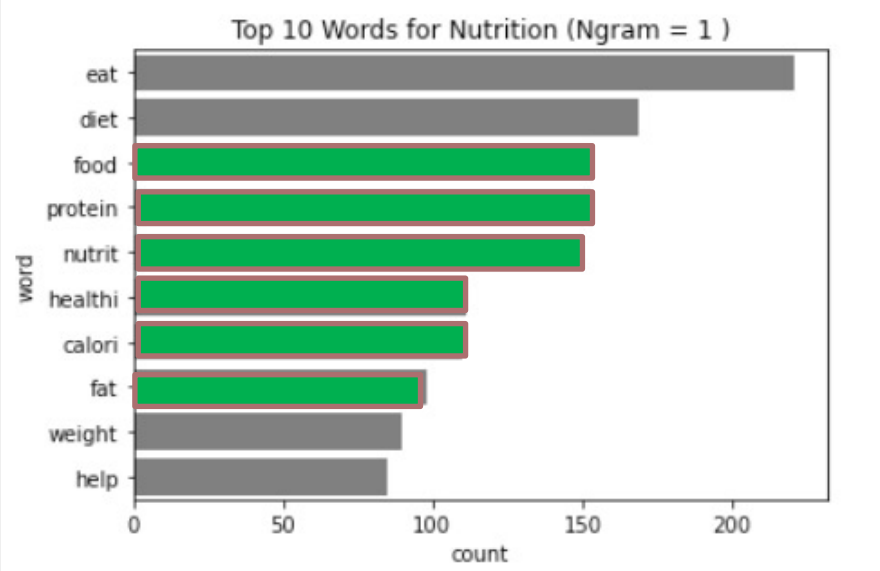
Word Analysis
Top 10 words

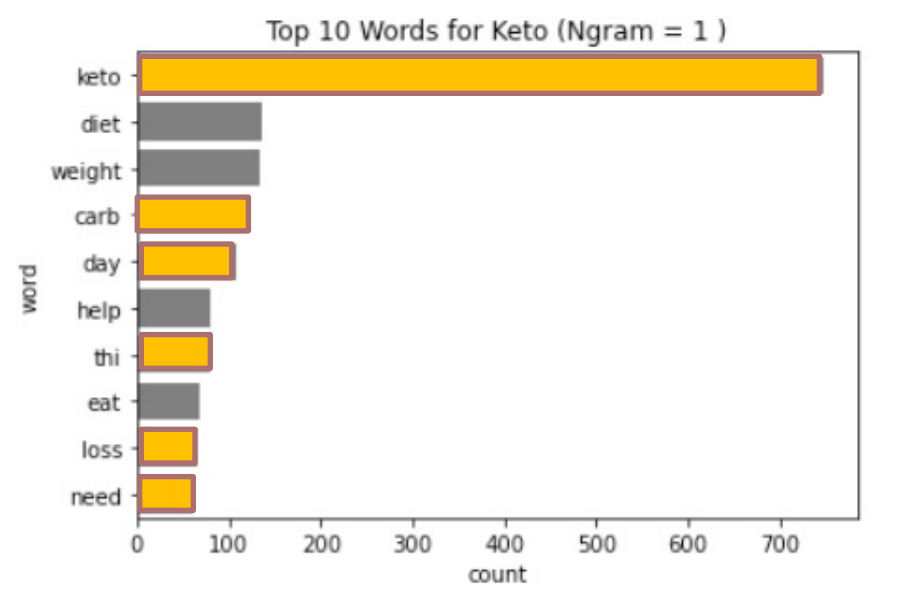




Similarities

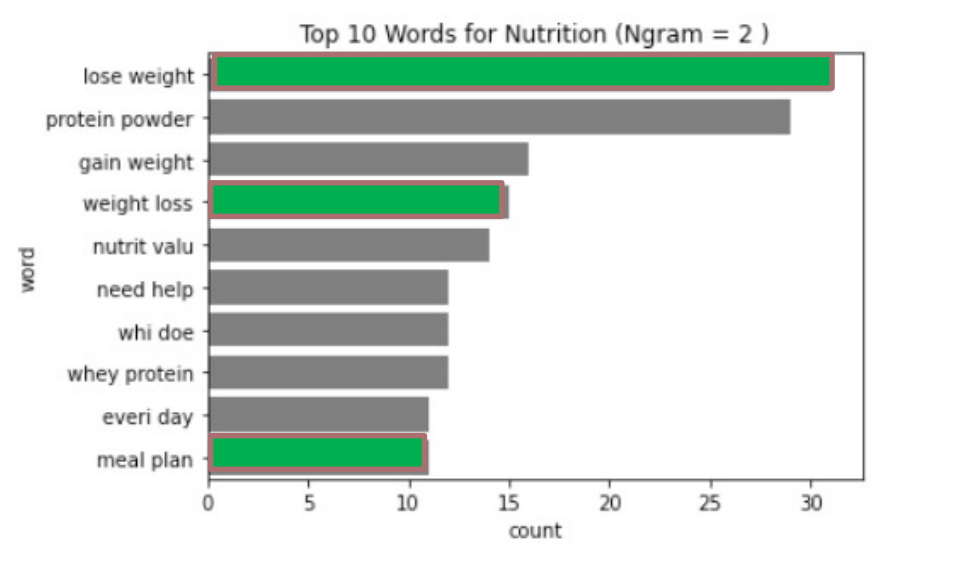
Word Analysis
Top 10 words

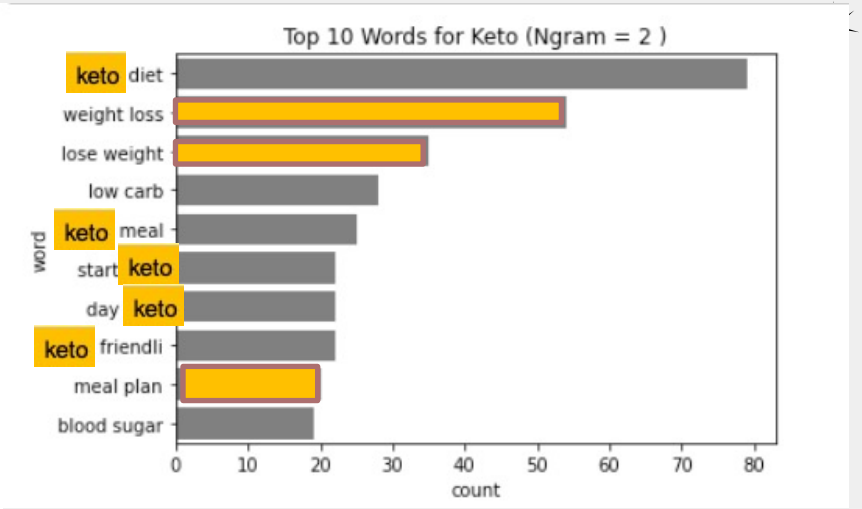




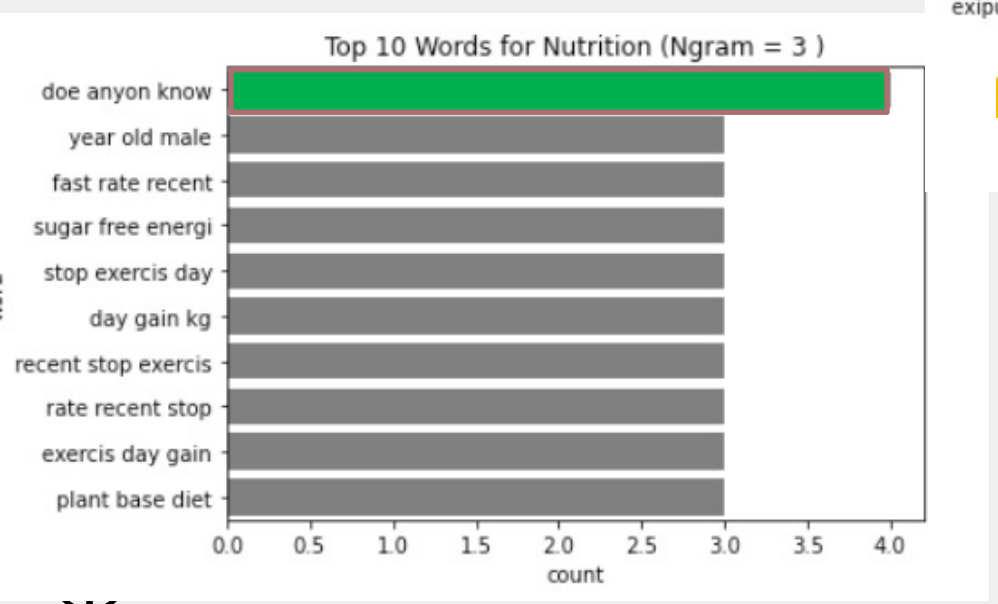
Differences

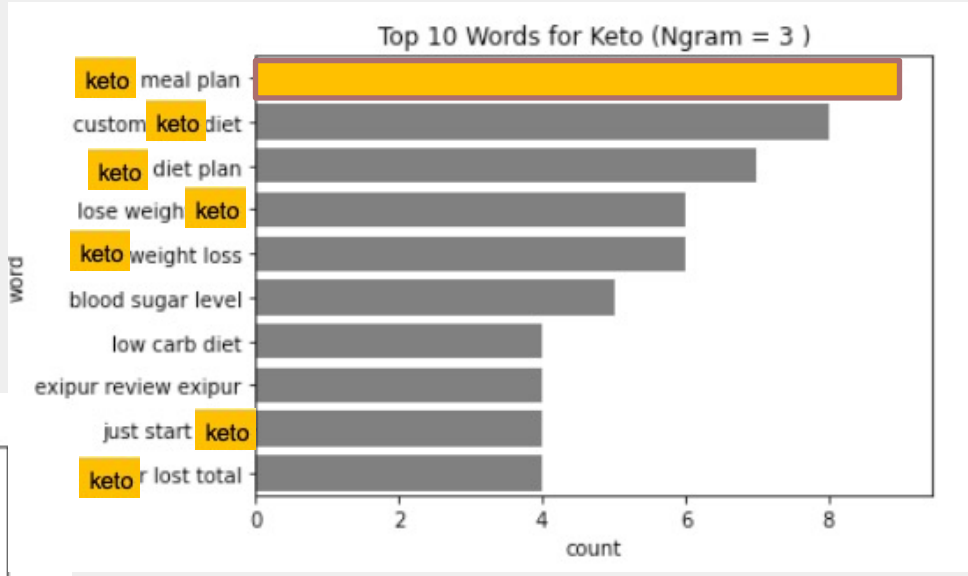
Word Analysis (BiGram)



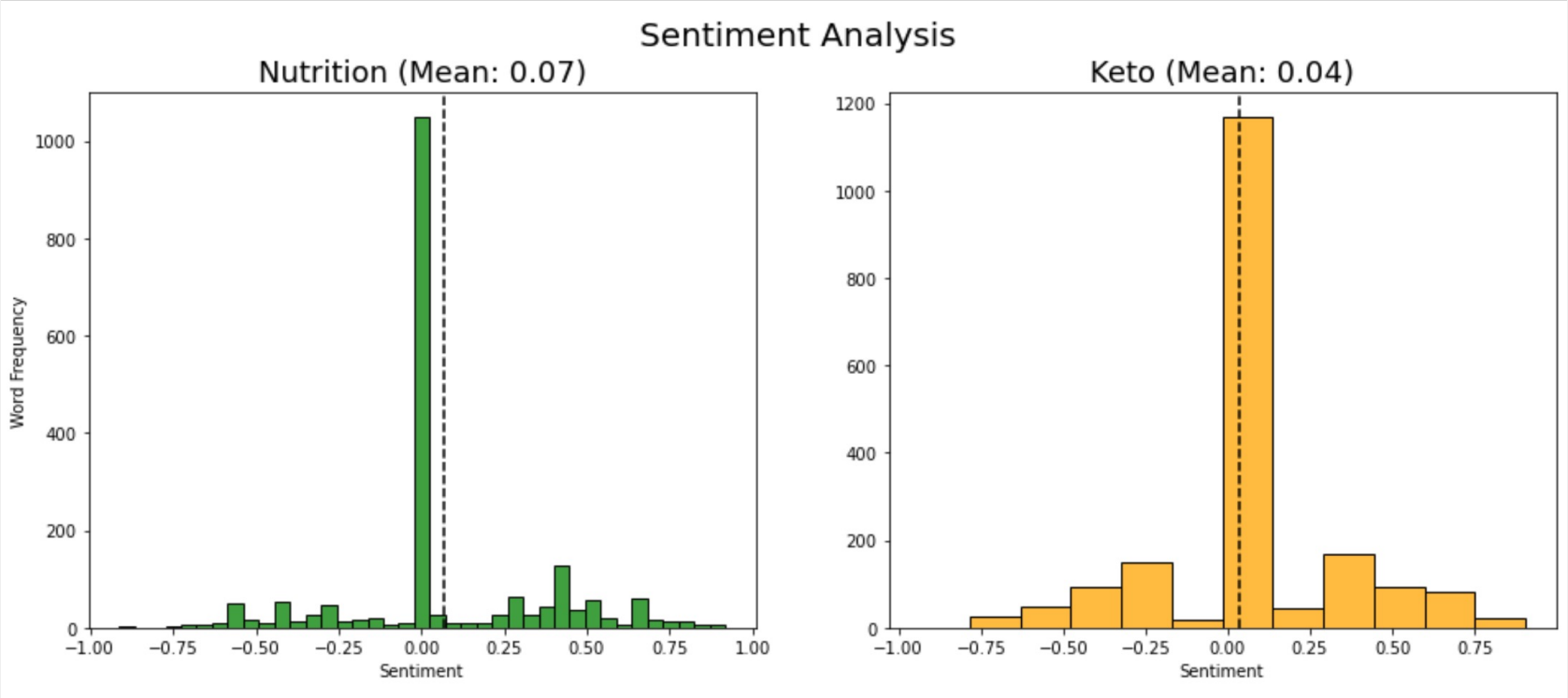


Word Analysis (TriGram)





Exploration / Sentiment Analysis



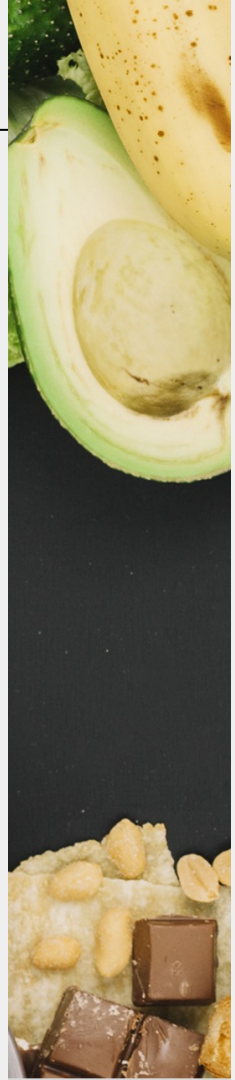
Nutrition has more positive sentiments than Keto

Nutrition

80%

*'What are some of the **best healthy** foods to eat to boost appetite and sustain energy?'*

'Best liver care for high alcohol consumption'



-70%

“I screwed up and ate 40 to 60 grams of saturated fat everyday for 2-3 years while bulking. Have I caused ireveisable damage to my body?”

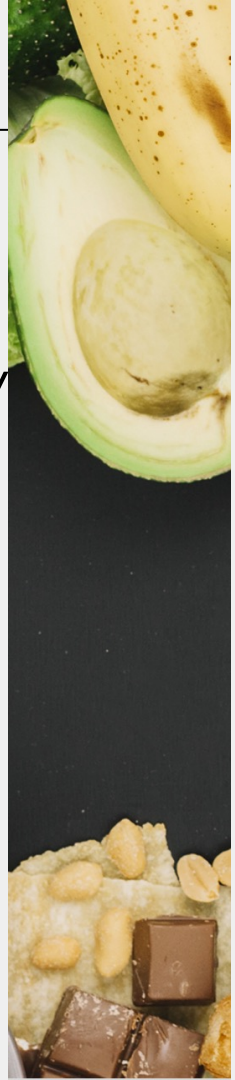
"I'm Scared My Diet Is Going To Kill Me Early"

Keto

80%

"That feeling of bliss when you cook a new recipe/dish for the first time and it's so good it brings a tear to your eye..."

"The best fish I've ever eaten!! What is your best food?"



-80%

'Do You Make These Four Common Diet Mistakes That Sabotage Your Health and Stall Fat Loss?'

'I want to stick to keto so bad, but I keep finding myself cheating'

MODEL SELECTION

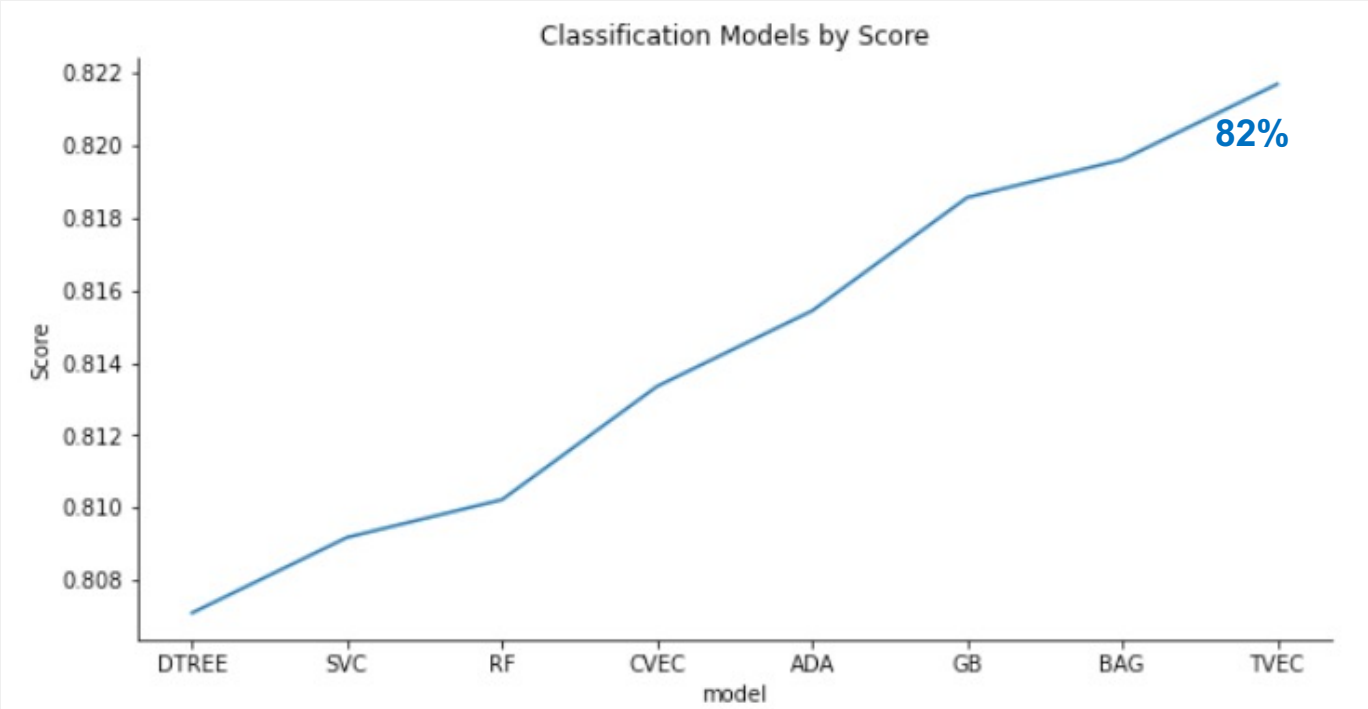
Our baseline model:

50%

CLASSIFICATION MODELS

8 Models:

DecisionTree
SVC
RandomForest
CountVectorizer
AdaBoost
GradientBoost
Bagging
TFIDF Vectorizer



Challenges:

1. Overfitting problem

2. Time to run GridSearch using hyperparameters:

Boosting is slowest amongst all models, as it builds the model sequentially instead of parallel like Random Forest and Bagging, so tuning the hyperparameters can take longer time.

TOP 2 Classification Models

TFIDF Vectorizer

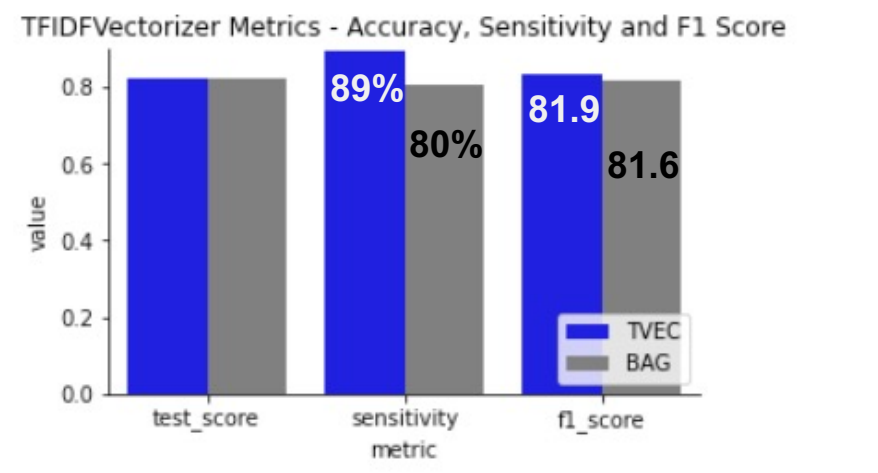
(1) Term Frequency (TF) - how many times a word appears in a document ***(2) Inverse document frequency (IDF)*** *the rarity of the word across a set of documents (corpus)*

Bagging Classifier

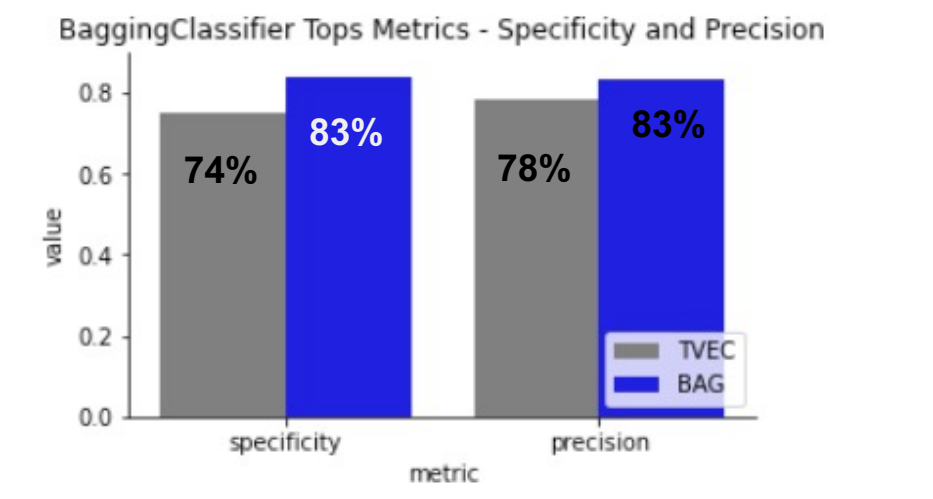
Ensemble Method – *The goal is to combine the predictions of several base estimators built with a given learning algorithm in order to improve generalizability / robustness over a single estimator.*

Bootstrap - *means random resampling of data and we control whether samples and features are drawn with or without replacement.*

How do we choose our “best” Model.. Metrics?

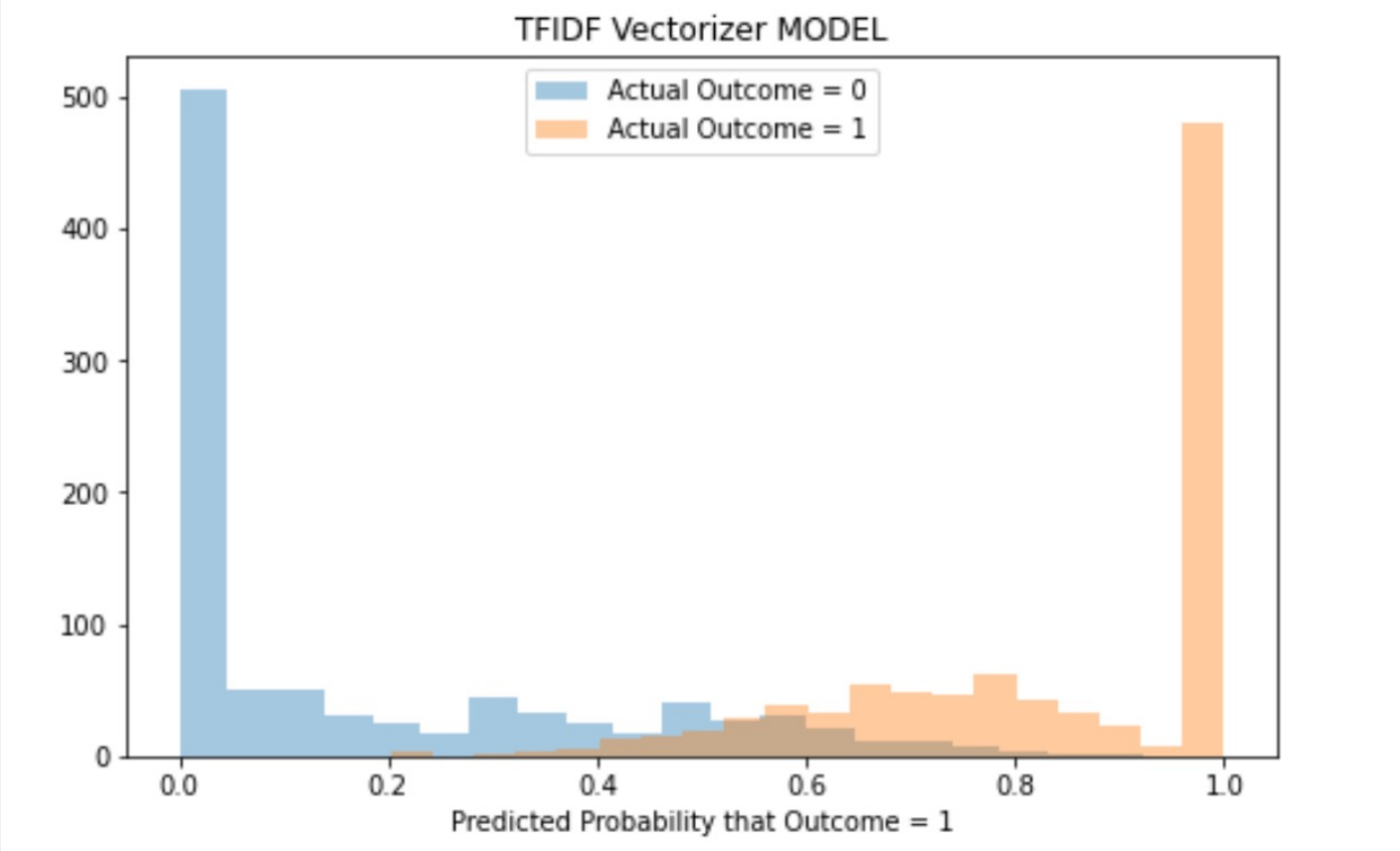


TFIDFVectorizer wins Accuracy, Sensitivity and F1 Score

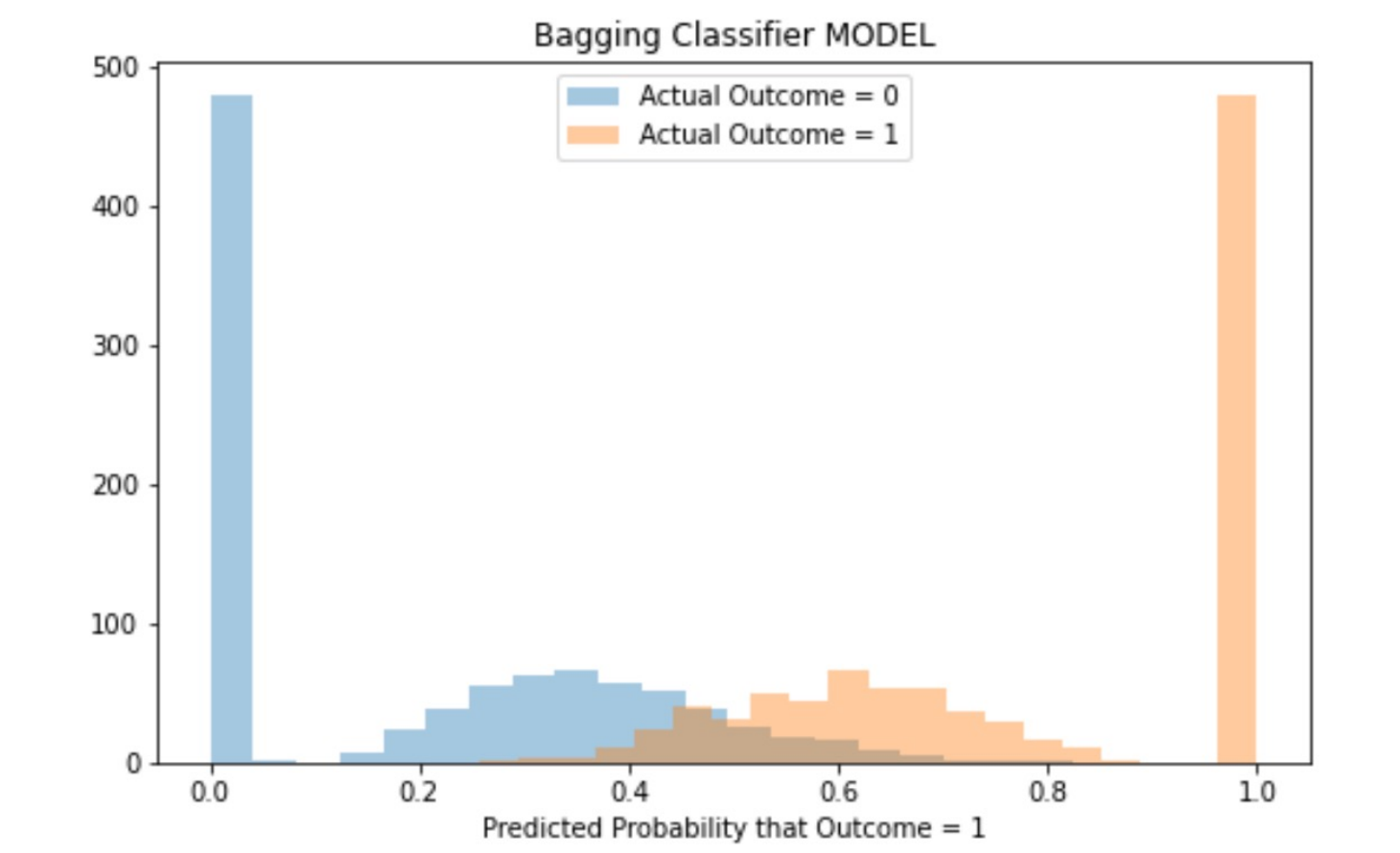


BaggingClassifier wins Specificity and Precision

Distribution of Probabilities – TFDIF Vectorizer



Distribution of Probabilities – Bagging Classifier



Selected Model: TFIDF Vectorizer

Reason: We selected ***TFIDF Vectorizer*** as our best model amongst all models for this classification project, as it is highest not just in Accuracy but also in F1-Score, as we want to balance Sensitivity and Specificity. It is also fast and easy to build with consistent result compared to ensemble methods like Bagging where data is randomly selected. In terms of interpretability, we can also look at the coefficients to validate the words that have highest correlation to each subreddits. Though there is overfitting, it is still not as bad as other models.

Best Params:

'best_params_': {'tvec__max_df': 0.2,
'tvec__max_features': 4000, 'tvec__min_df': 1,
'tvec__ngram_range': (1, 2),
'tvec__stop_words': None},

Model Insight: Misclassified Post

False Negative

	title	actual	predict
	FREE KETO COOKBOOK	Nutrition	Keto
	Do keto diet work?	Nutrition	Keto
	Best Keto Plan recipes	Nutrition	Keto

Model Insight: Misclassified Post

False Positive

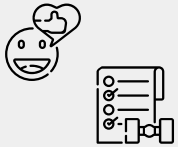
	title	actual	predict
	Vitamin c	Keto	Nutrition
	Protein powders?	Keto	Nutrition
	Inaccurate nutrition facts	Keto	Nutrition
	Great Value chia seeds nutritional info vs other sources of nutritional info	Keto	Nutrition
	For the purpose of blood sugar/insulin regulation, how important is it to eat protein paired with fat/fiber?	Keto	Nutrition
	Floralite is a dietary supplement containing units of pure, alive and active microbes designed to nourish your gut, making you lose fat incredibly fast.	Keto	Nutrition
	Wicked Protein	Keto	Nutrition
	eat healthy	Keto	Nutrition
	Vitamin Supplements	Keto	Nutrition

To answer our Business Problem:

Yes, we have created a TFDIF Vectorizer model that can classify posts from two different subreddits based on the TITLW with 82% Accuracy.

Recommendation:

Tuning of hyperparameters to overcome overfitting



Include other features like self-text and probably sentiment analysis score might improve our metrics.

Include images or videos in our analysis for more accurate prediction (which requires more knowledge on different ML domains)

A PICTURE IS WORTH A THOUSAND WORDS -- #WORDCLOUD


slidesgo

Thank you!

Rebellon Carina

Data Analyst | Data Engineer |
Aspiring Data Scientist