# Nutrition vs Keto

*Subreddit Classification*



# Data Science Process

**01** Define Business problem.

**02** Gather data.

**03** EDA , Cleanup & Preprocessing

**04** Build Models & Evaluate Metrics.

**05** Answer Business problem.

**06** Recommendation

# 01. Business Problem:

**We want to help the moderators to automate the classification and this will benefit the maintenance and sanity of each subreddits by posting to the correct group.**

**Can we accurately label if a post is more relevant for Nutrition or Keto?**

# 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.

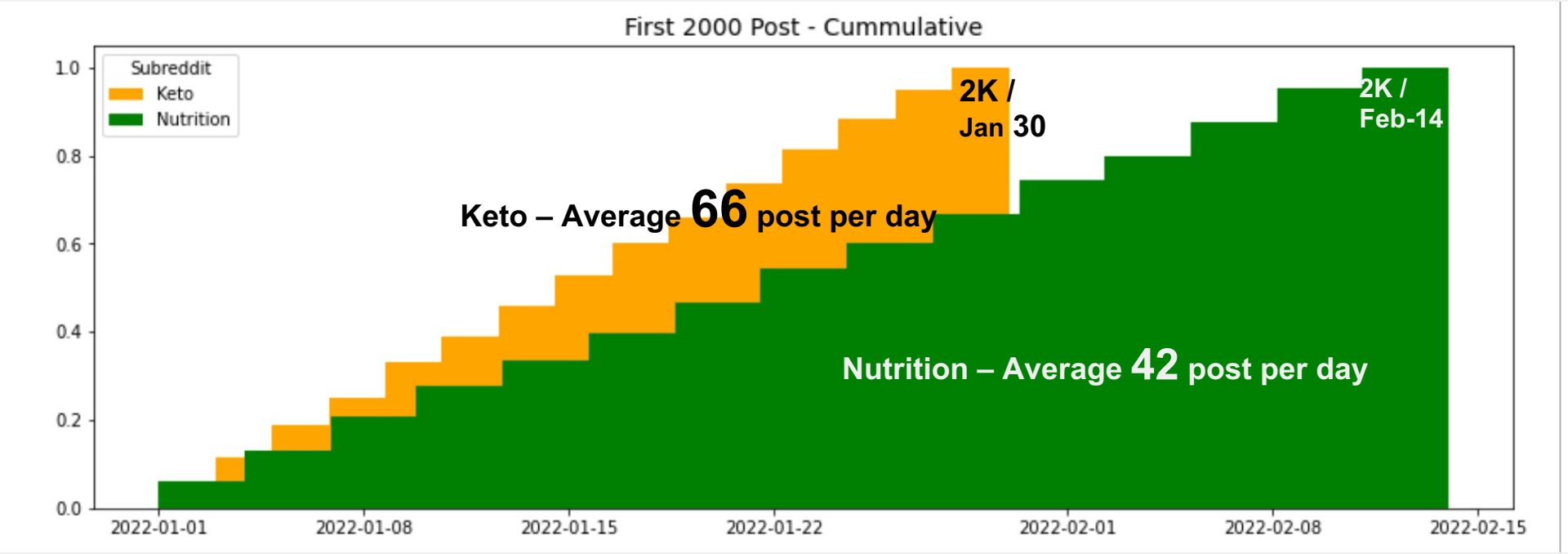
**2.3M Members**

## 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**

# 02. Gather Data (Data from Jan to Feb '22)



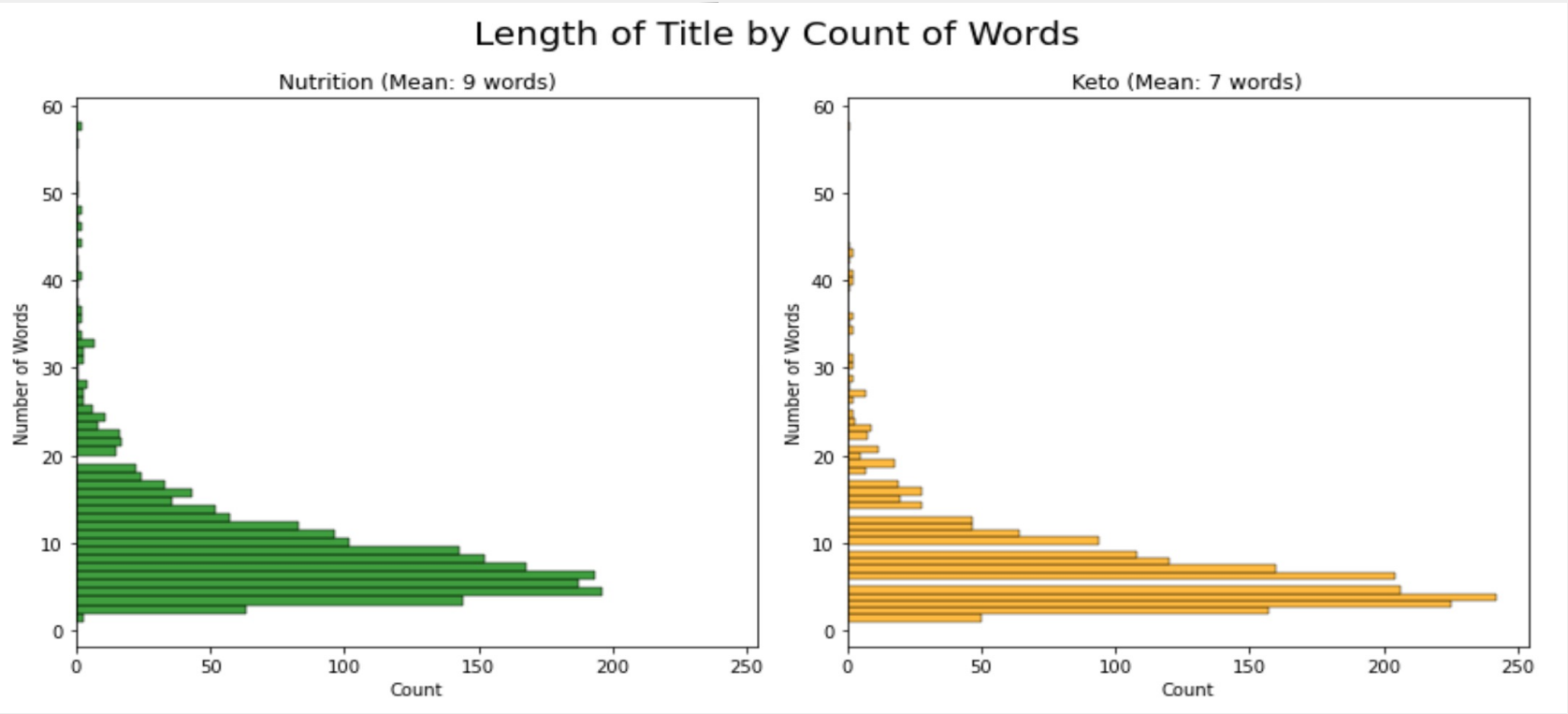

***First 2000 post from Jan 1, 2022:***

**Since Keto is more active and have more members, it reached the 2K post on Jan 30, 2022. Nutrition's reached 2K post on Feb 14, 2022.**

# 03. EDA
# Feature(s)?

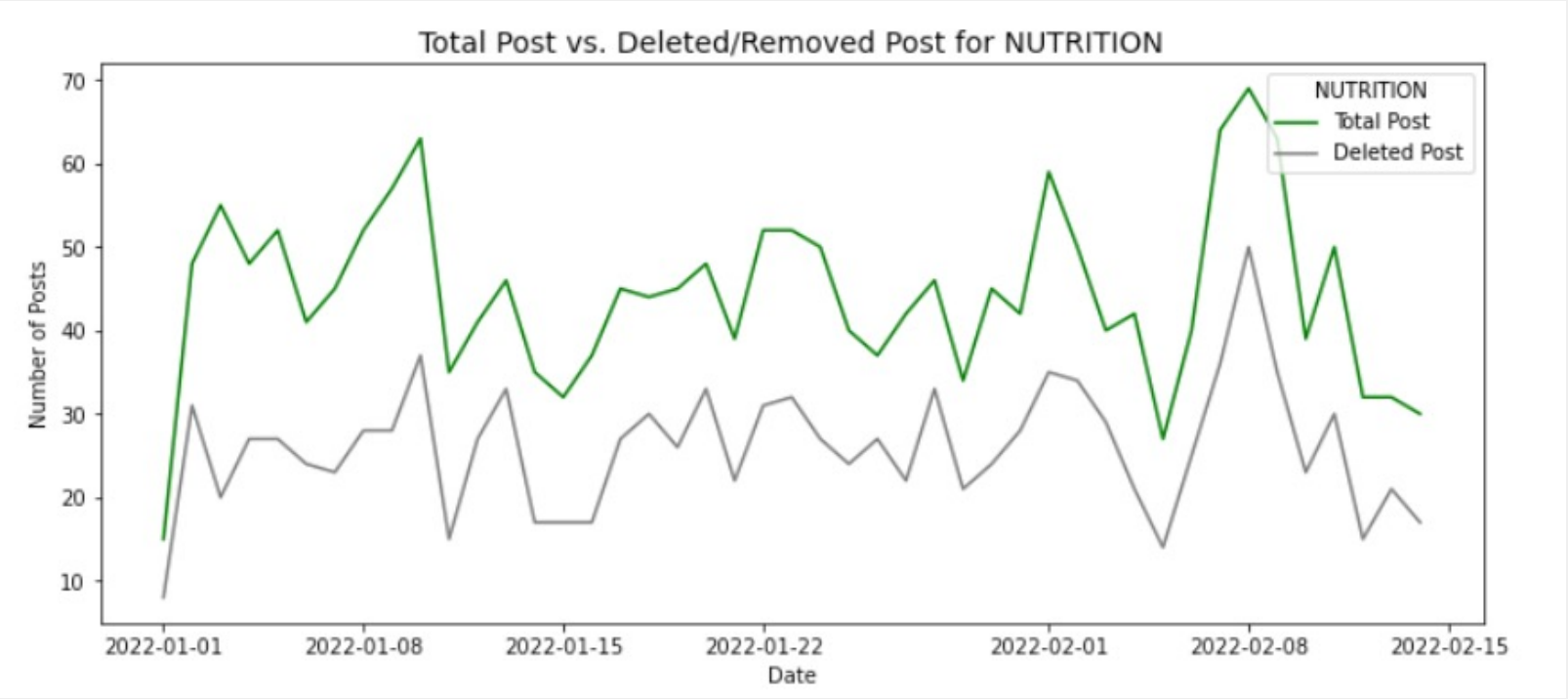
## Title or Self-text,
## ... Or both?

# Title Analysis



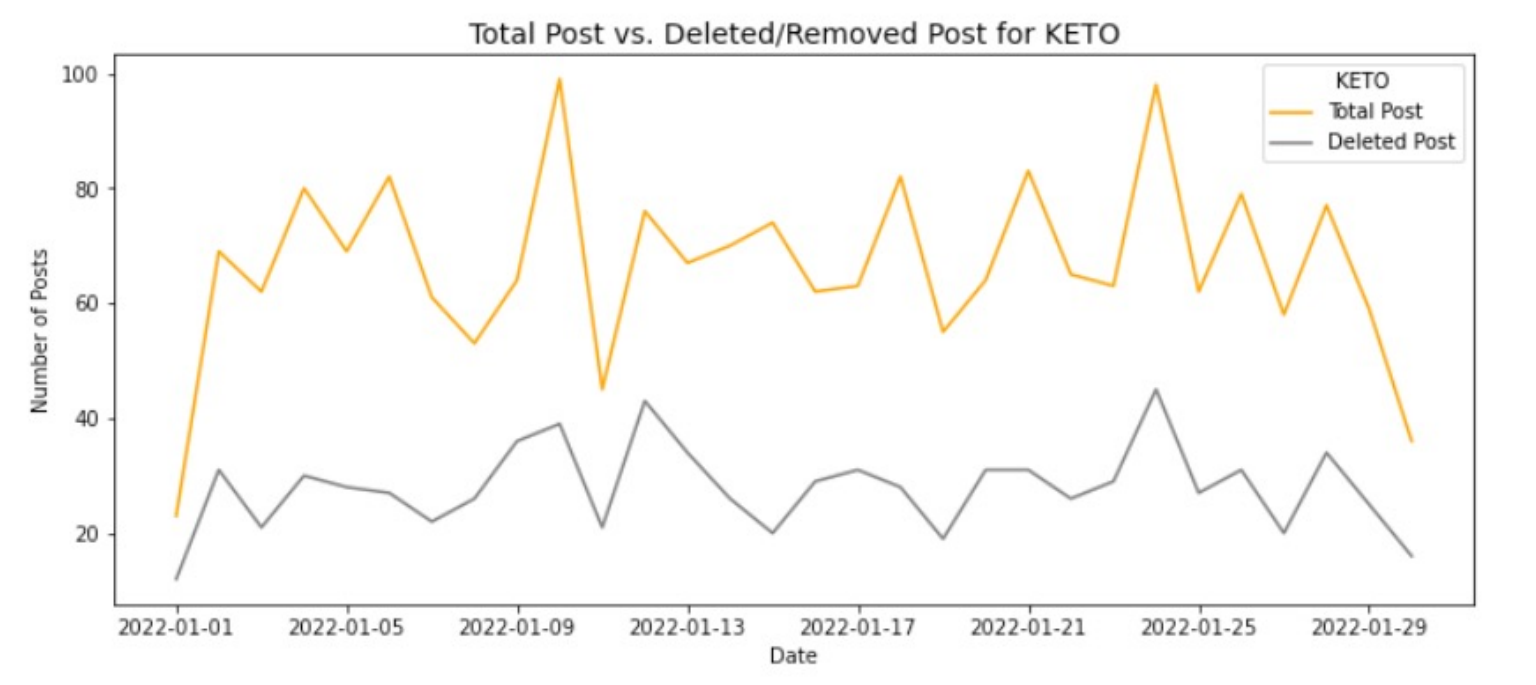

**Nutrition has slightly higher average count of words, they have (almost) similar distributions**

# Self-Text Analysis - Nutrition




**59% of the self-text was removed/deleted**

# Self-Text Analysis - KETO




**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)

# Cleanup and Pre-processing

***Cleanup Author:***

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

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

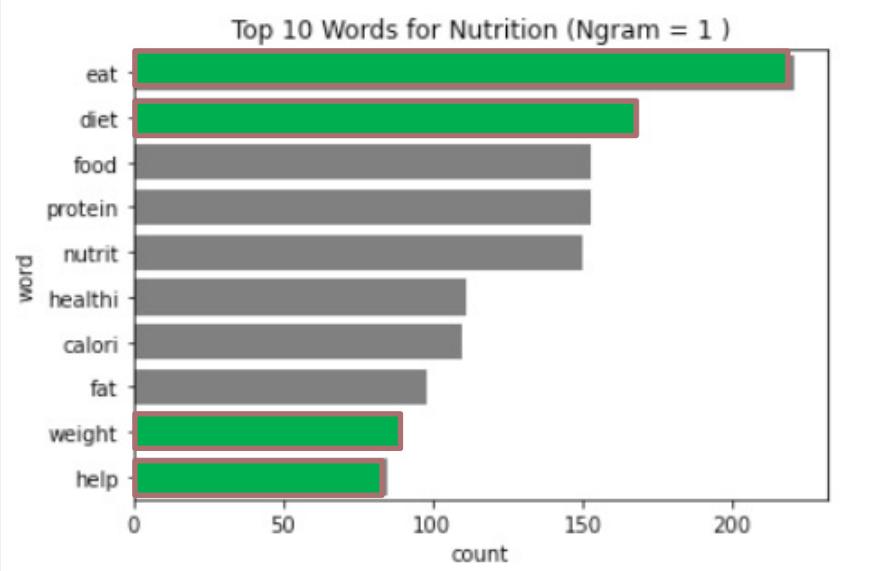
***Remove Duplicate:***

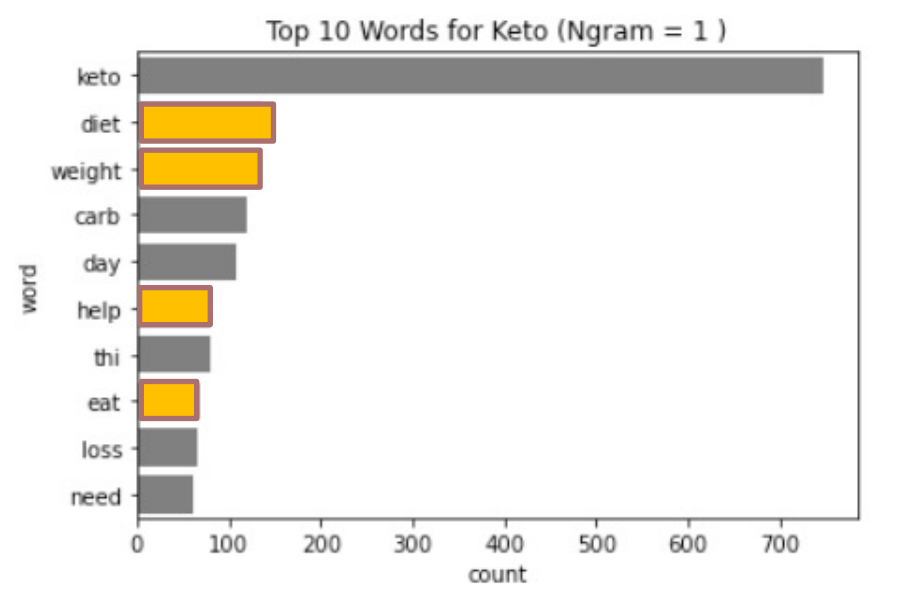
***Title*** **2.0%**

- **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*)

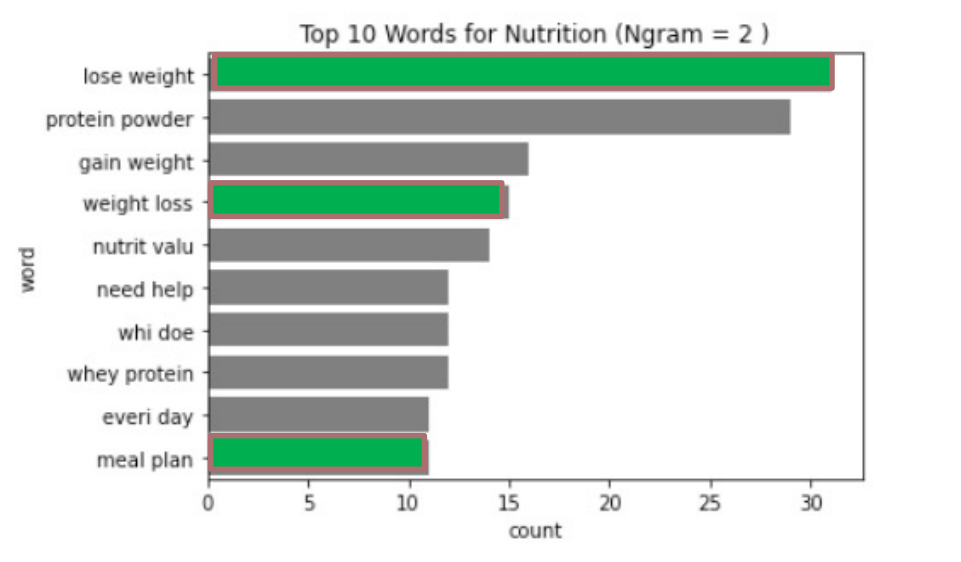


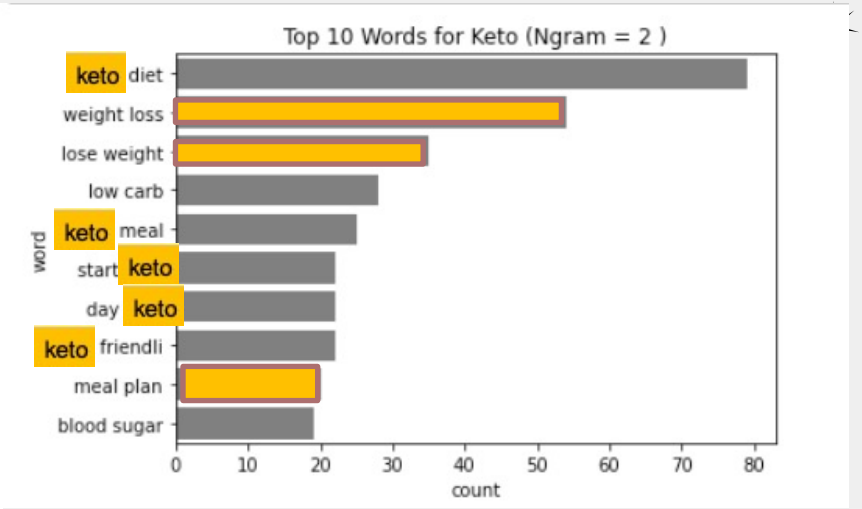
# Word Analysis
# Top 10 words



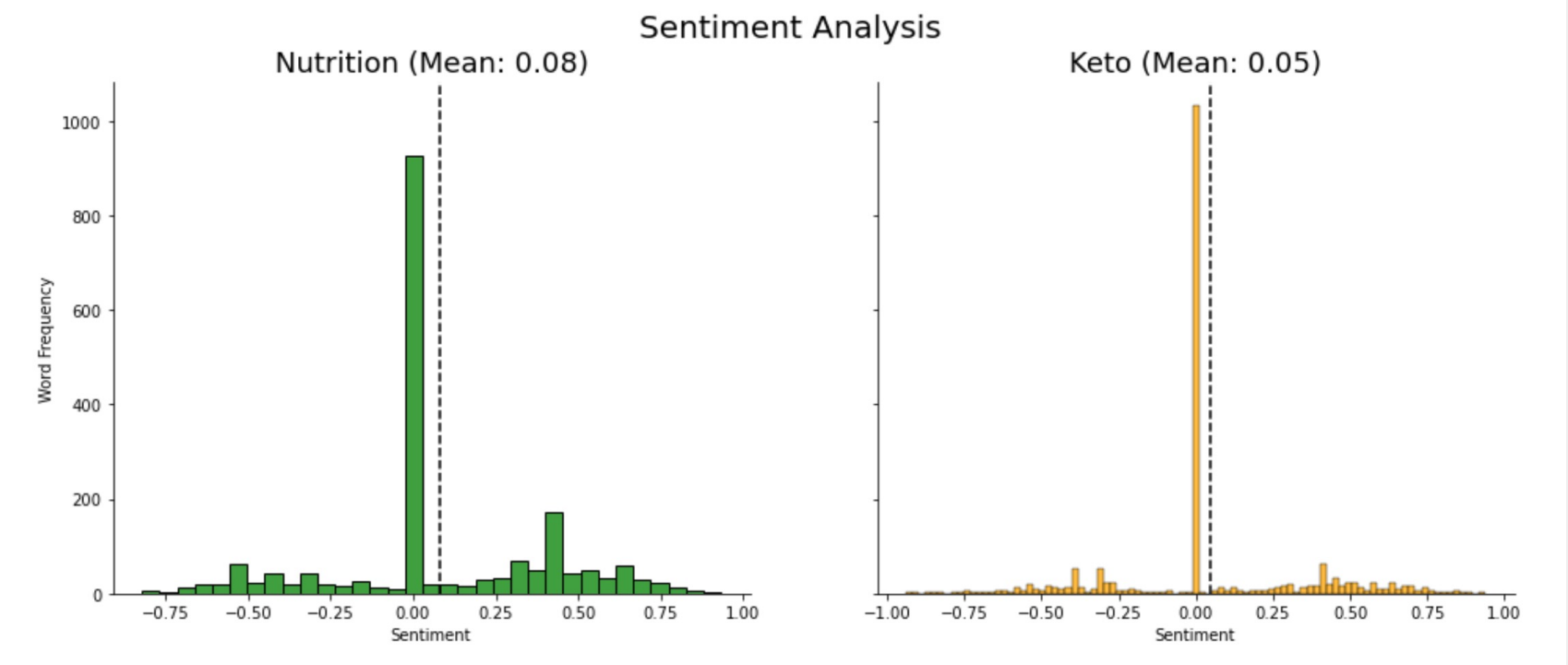



# Word Analysis (BiGram)



# Sentiment Analysis using VADER




Keto has higher “Neutral”

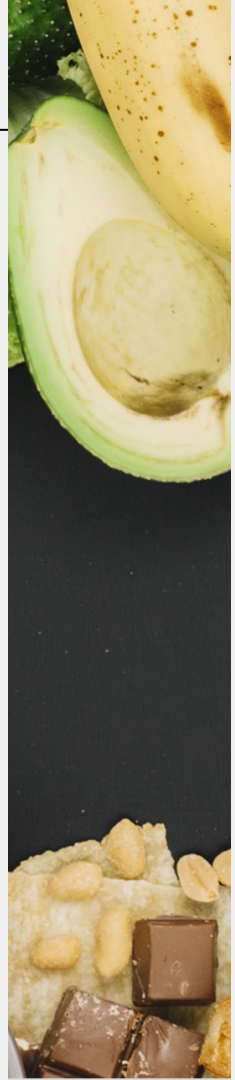
Nutrition has more positive sentiments than Keto.

# 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?”*

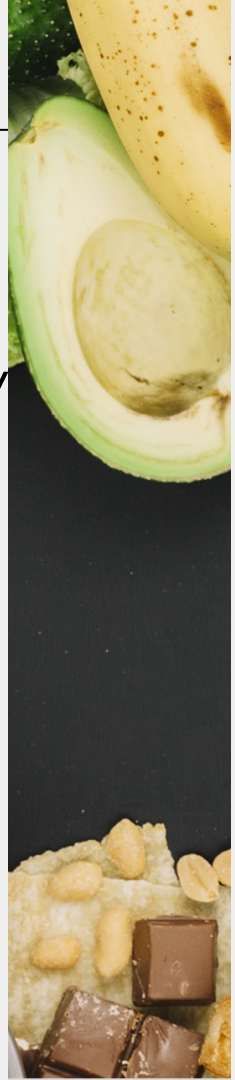
*"I'm Scared My Diet Is Going To Kill Me Early"*

# 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'*

# 04. Model with data & Evaluate model.

## Our baseline model:

**50%**

## Challenges:

### *1. Overfitting problem*

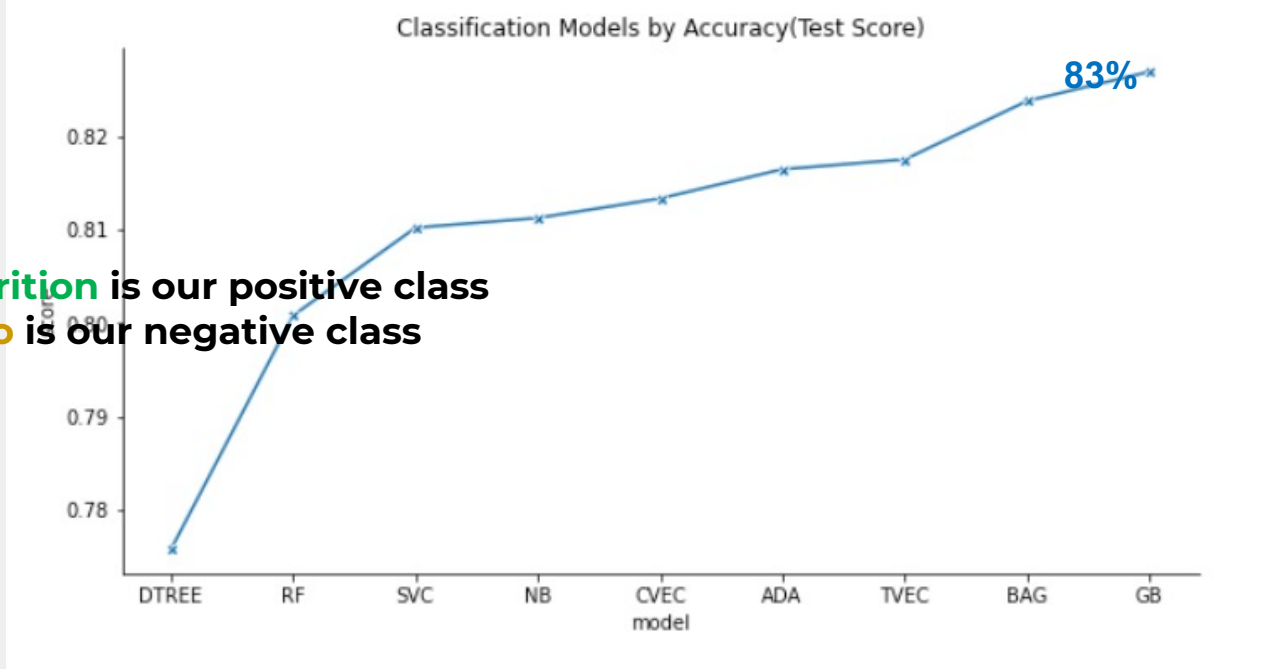

**2. Time to run GridSearch using hyperparameters:** *Boosting is slowest amongst all models, as it builds the model in sequential way (compared to Bagging where it trains in parallel or independently) so tuning the hyperparameters can take longer time for Boosting*

# MODEL's Accuracy

***9 Classification Models:***

**DecisionTree**
**RandomForest**
**SVC**
**NaiveBayes**
**CountVectorizer**
**AdaBoost**
**TFIDF Vectorizer**
**BaggingClassifier**
**GradientBoost**

**Nutrition is our positive class**
**Keto is our negative class**



# Classification Models – Ensemble Technique

## GradientBoost Classifier

***Boosting*** *is a method of converting weak learners into strong learners. Boosting takes a weak base learner and tries to make it a strong learner by retraining it on the misclassified samples.*

*In* ***gradient boosting****, it trains many model sequentially. Each new model gradually minimizes the loss function of the whole system using Gradient Descent method.*

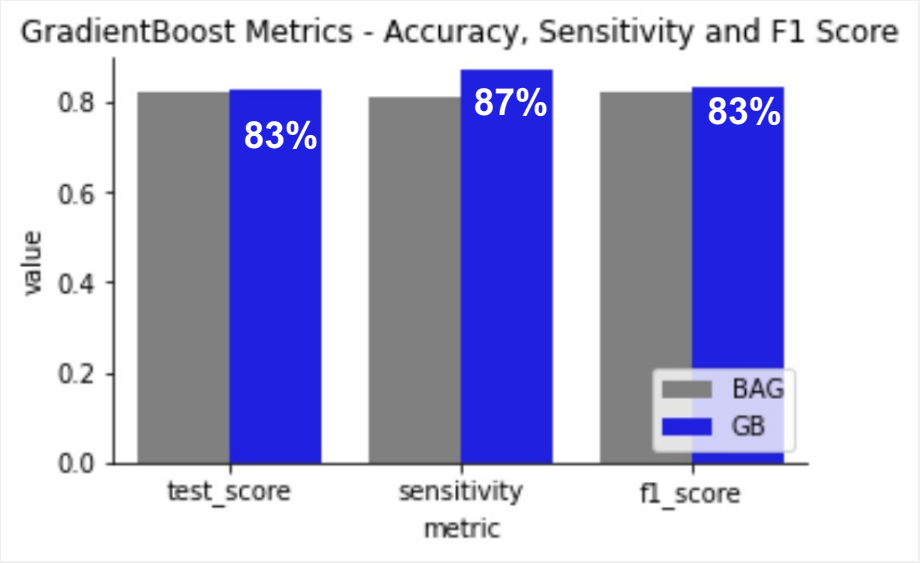
## Bagging Classifier

***Bagging*** - single training algorithm is used but on different subsets of the training data. Subsets are bootstrapped data where random data is generated with replacement.

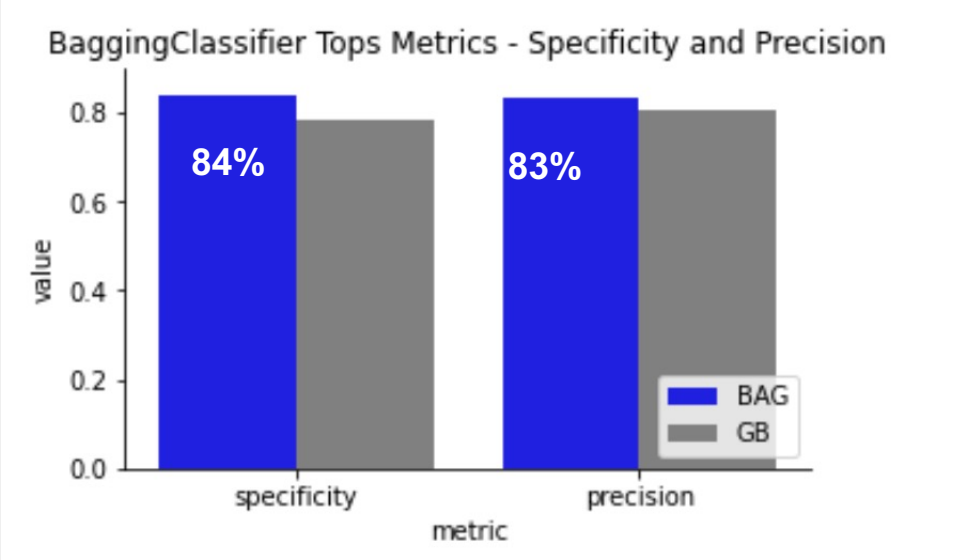
For aggregating the outputs of base learners, bagging uses majority *voting for classification*

# Other Metrics :

**Nutrition** is our positive class
**Keto** is our negative class



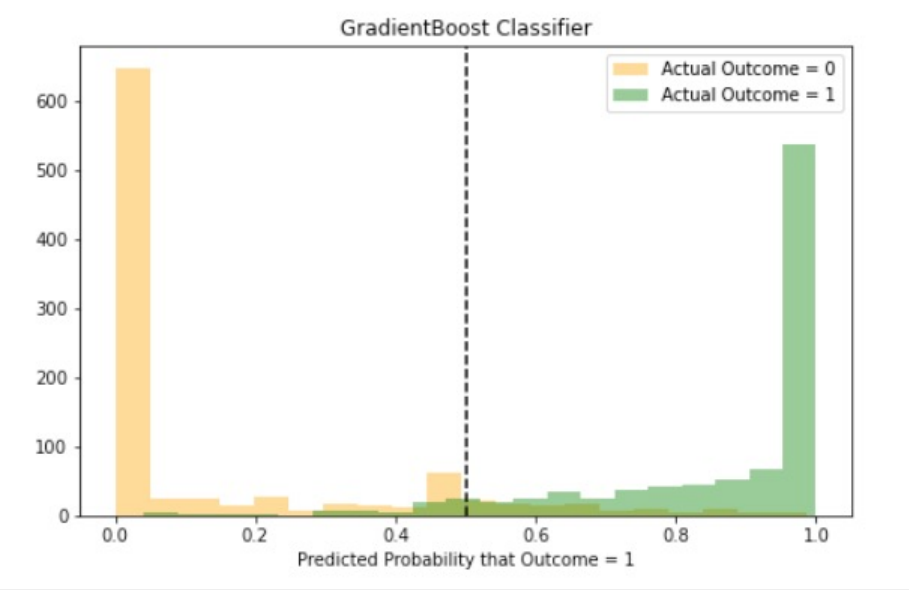

**GradientBoost wins Accuracy, Sensitivity and F1 Score**

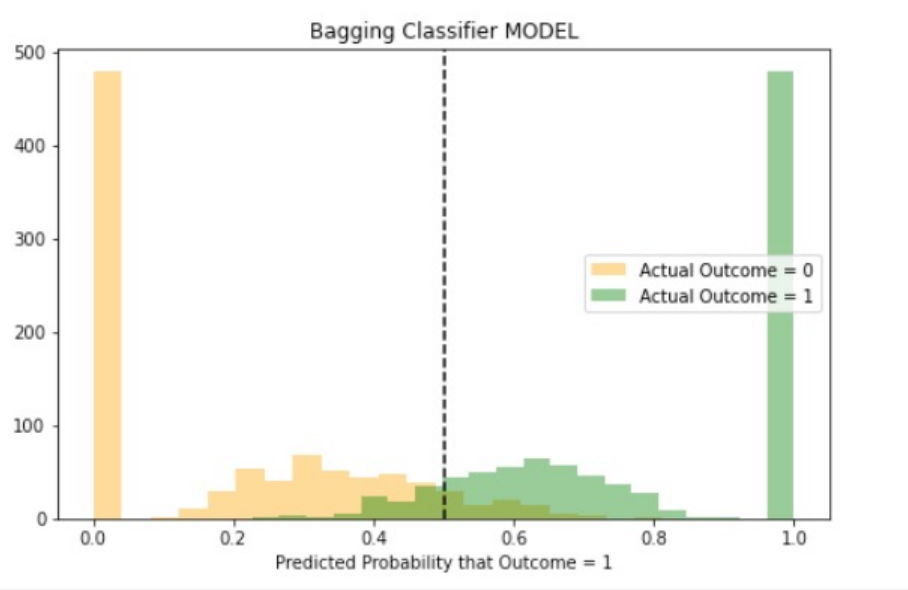



**BaggingClassifier wins Specificity and Precision**

# Distribution of Probabilities

**Nutrition** is our positive class
**Keto** is our negative class



# Selected Model: *Bagging Classifier*

**Reason:** We selected ***Bagging Classifier*** as our best model for this classification project, though GradientBoost Classifier tops Accuracy and F1, it wasn't that much difference. The major difference is in Sensitivity and Specificity, we want our negative class to be predicted more accurately, as **Keto** is strict diet and should not be classified under the **Nutrition** subreddit. Bagging Classifier is highest in ***Specificity*** (True Negative) rate with **83.5%** rate compared to GradientBoost Classifier with only **78.5%**

| Metric: | Bagging Classifier Metric % | Difference from Gradient Boosting |
|---|---|---|
| Test Score (Accuracy) | 82.30 % | -0.31 |
| Sensitivity | 81.25 % | -5.63 |
| Specificity | 83.50 % | +5.00 |
| Precision | 83.16 % | +2.96 |
| F1 Score | 82.19 % | -1.21 |

# Model Insight: Misclassified Post (Bagging Classifier Model)

## False Negative

| title | actual | predict |
|---|---|---|
| the fastest method to lose weight https://shrinke.me/lqwFdwjU | Nutrition | Keto |
| New member i need halp | Nutrition | Keto |
| 36F with PCOS, trying Low carb ntermittent fasting - so tired, how can I fix this? | Nutrition | Keto |
| Keto Diet and its health benefits | Nutrition | Keto |

## False Positive

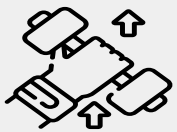
| title | actual | predict |
|---|---|---|
| Great Value chia seeds nutritional info vs other sources of nutrition info | Keto | Nutrition |
| For the purpose of blood sugar/insulin regulation, how important is it to eat protein aired with fat/fiber? | Keto | Nutrition |
| I'm making chicken teriyaki tomorrow, can someone recommend a side that will complement it well? | Keto | Nutrition |
| How many calories ould I be consuming daily? | Keto | Nutrition |

# 05. To answer our Business Problem:

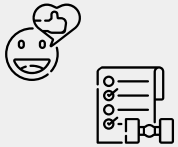
***Yes, we have selected Bagging Classifier as our model that can label TITLE as Keto or Nutrition with 82% Accuracy.***



# 06. 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