# Project 3: Web APIs & NLP

Group 2
Jerome, Kah Beng, Miao Er

# Background

Keto Diet: High-protein, low-carbohydrate, and fat-rich diet.

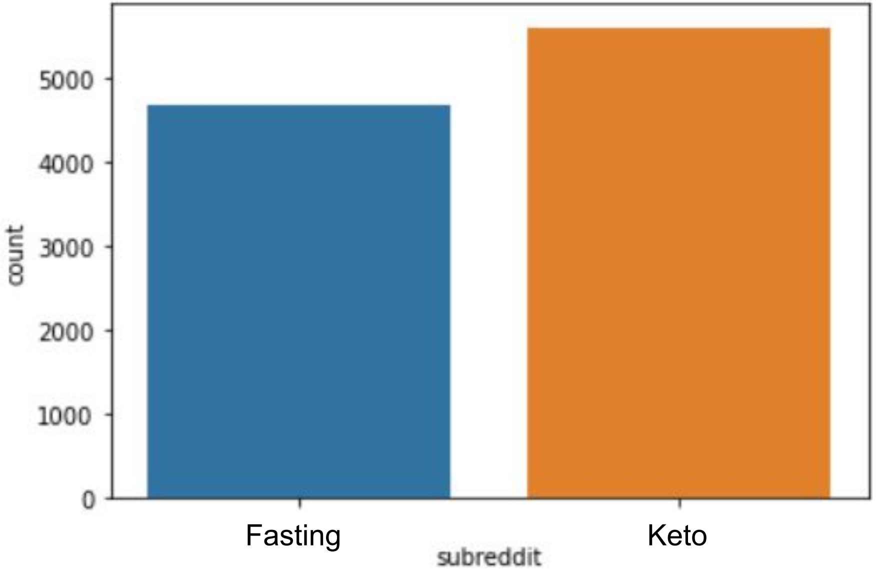
Intermittent Fasting: eating pattern that involves extended period of time, typically 14 hours or longer, are not eating.

# Problem Statement

- Data scientists at a startup firm being engaged by a meal delivery services company focusing on intermittent fasting diet meal plans
- To assist the company in their social media marketing/advertising campaigns by building a classification model to understand the current popular/trending words associated with intermittent fasting and keto
- To understand the current sentiment analysis of the public on these 2 diet trends
- To provide secondary stakeholder information on the diets from sentiment analysis

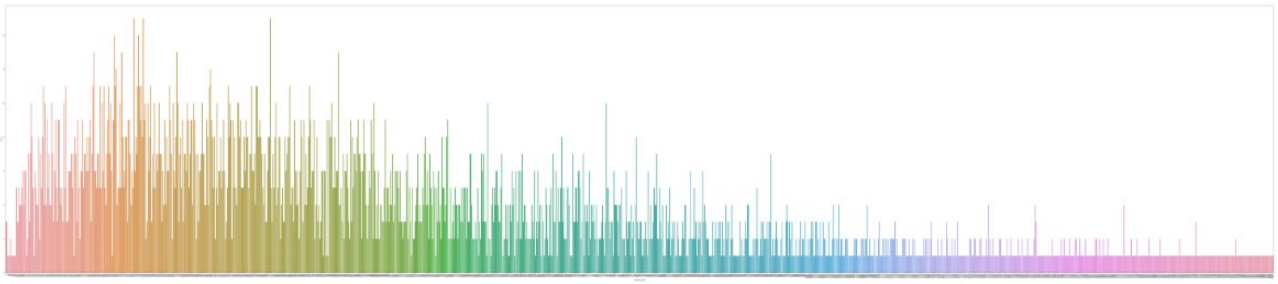
# Data Wrangling

- Reddit's Pushshift API
- >10,000 posts
- Intermittent fasting: 4599, Keto diet: 5844

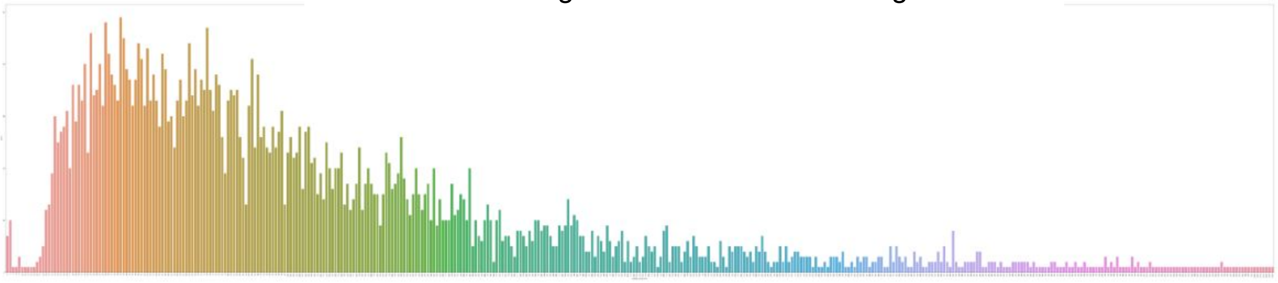


# EDA & Visualisation

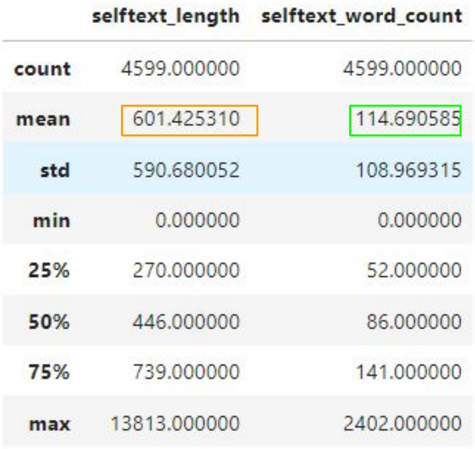
# EDA on Selftext - Intermittent Fasting



Intermittent Fasting - Selftext Character Length



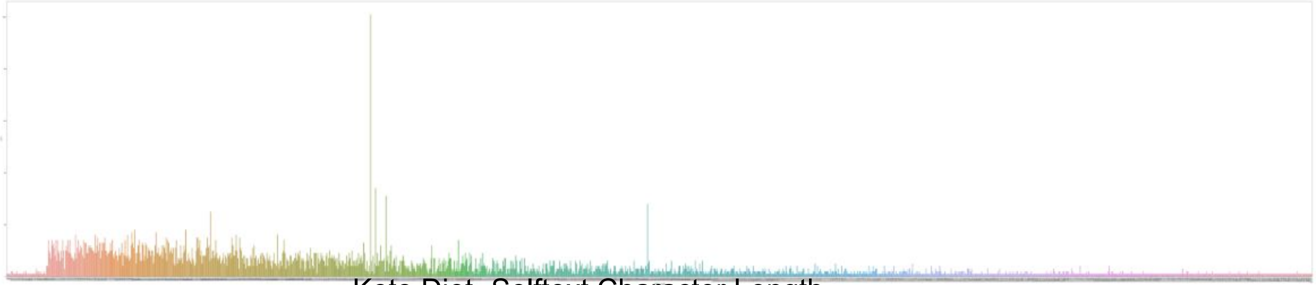
Intermittent Fasting - Selftext Word Count



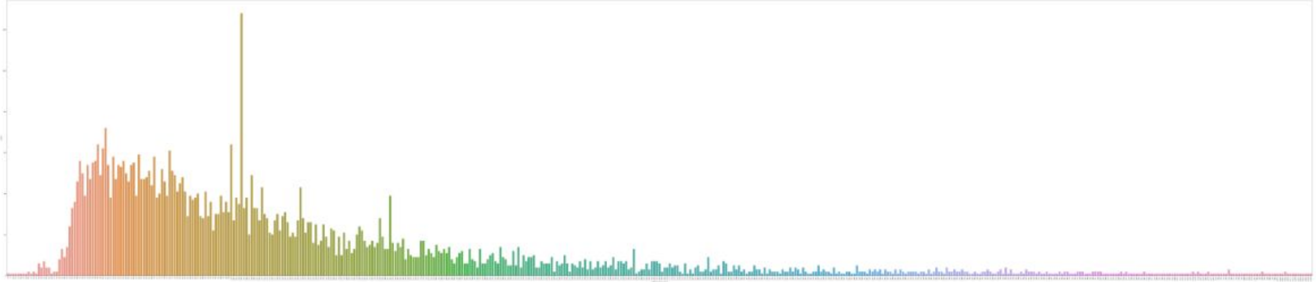
| | selftext_length | selftext_word_count |
|---|---|---|
| **count** | 4599.000000 | 4599.000000 |
| **mean** | 601.425310 | 114.690585 |
| **std** | 590.680052 | 108.969315 |
| **min** | 0.000000 | 0.000000 |
| **25%** | 270.000000 | 52.000000 |
| **50%** | 446.000000 | 86.000000 |
| **75%** | 739.000000 | 141.000000 |
| **max** | 13813.000000 | 2402.000000 |

- Different in mean of character length and word count
- Value count are similar in countplot shown

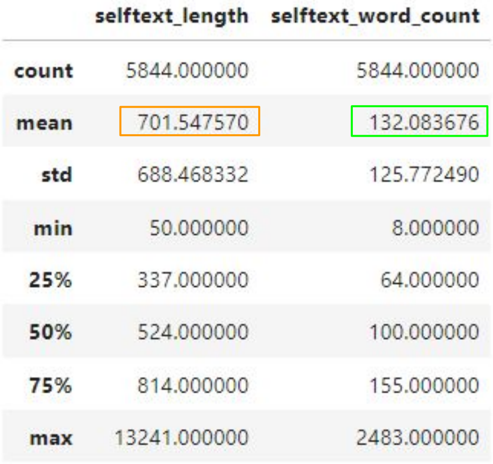
# EDA on Selftext - Keto Diet



Keto Diet- Selftext Character Length



Keto Diet- Selftext Word Count

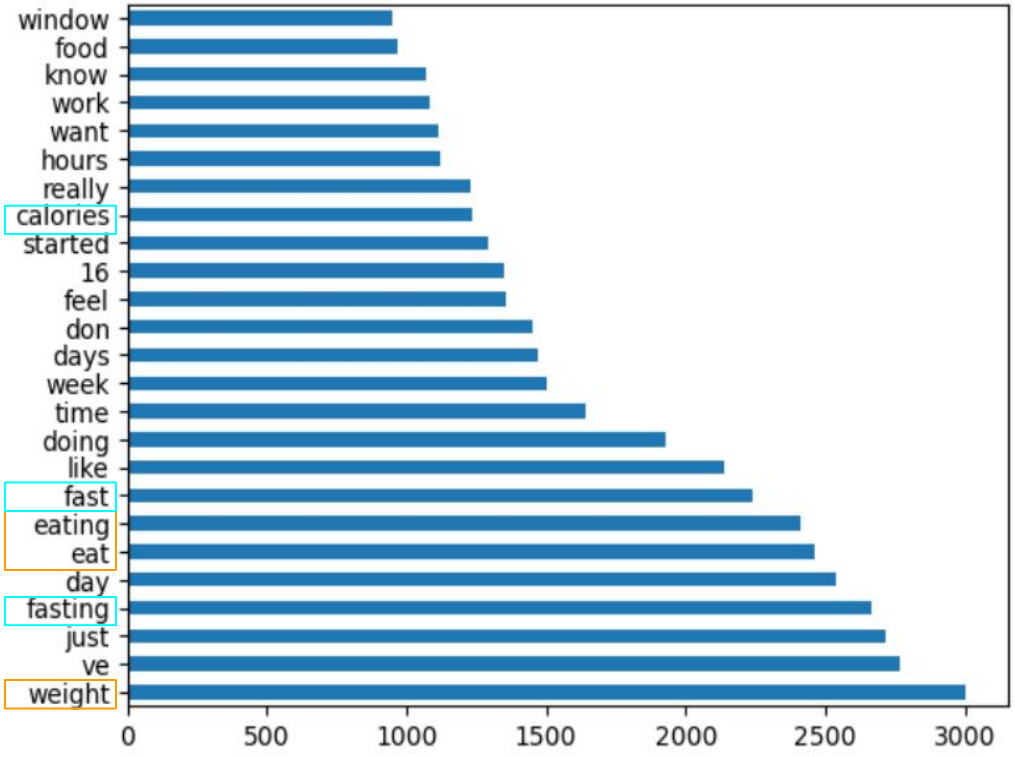


| | selftext_length | selftext_word_count |
|---|---|---|
| **count** | 5844.000000 | 5844.000000 |
| **mean** | 701.547570 | 132.083676 |
| **std** | 688.468332 | 125.772490 |
| **min** | 50.000000 | 8.000000 |
| **25%** | 337.000000 | 64.000000 |
| **50%** | 524.000000 | 100.000000 |
| **75%** | 814.000000 | 155.000000 |
| **max** | 13241.000000 | 2483.000000 |

- Different in mean of character length and word count
- Value count are similar between character length and word count as countplot shown
- Keto has higher character length and word count

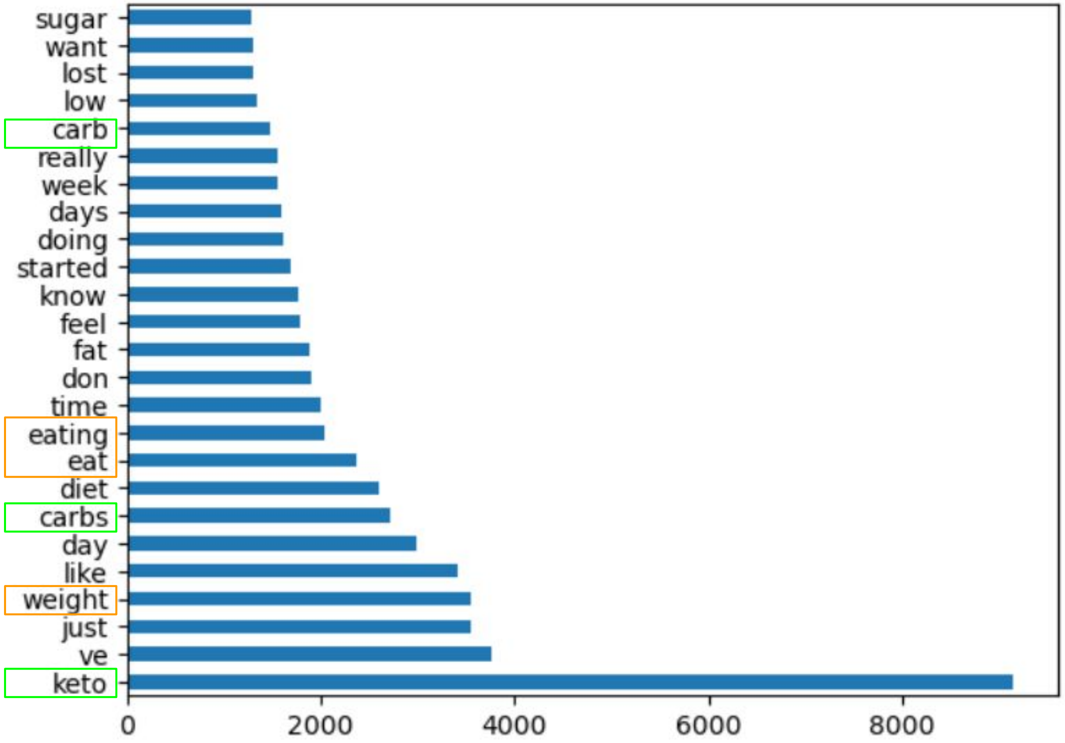
# EDA on word counts

## Frequency of the most common words – Intermittent fasting subreddit



- Overlap common words: weight, eat, eating, etc.
- Only present in intermittent fasting: fasting, fast, calories

# Frequency of the most common words – Keto subreddit

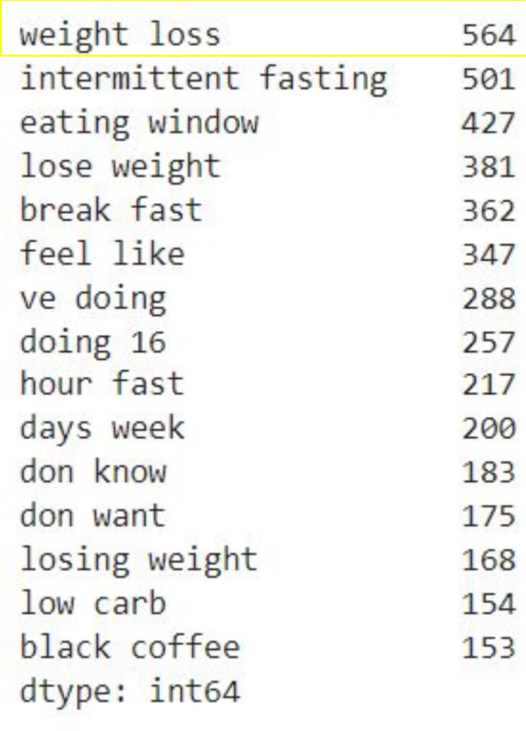


- Overlap common words: weight, eat, eating, etc.
- Only present in keto: keto, carbs, diet

**Based on the differences, we are able to build classification model for both subreddits.**

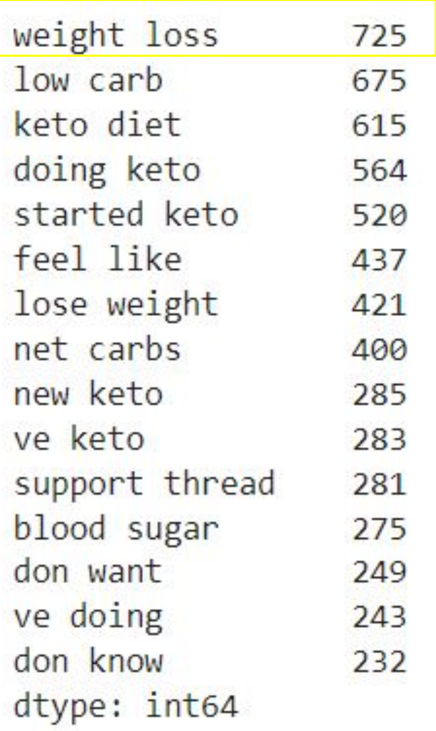
# Most common bigrams

## Intermittent fasting subreddit



```
weight loss              564
intermittent fasting     501
eating window            427
lose weight              381
break fast               362
feel like                347
ve doing                 288
doing 16                 257
hour fast                217
days week                200
don know                 183
don want                 175
losing weight            168
low carb                 154
black coffee             153
dtype: int64
```

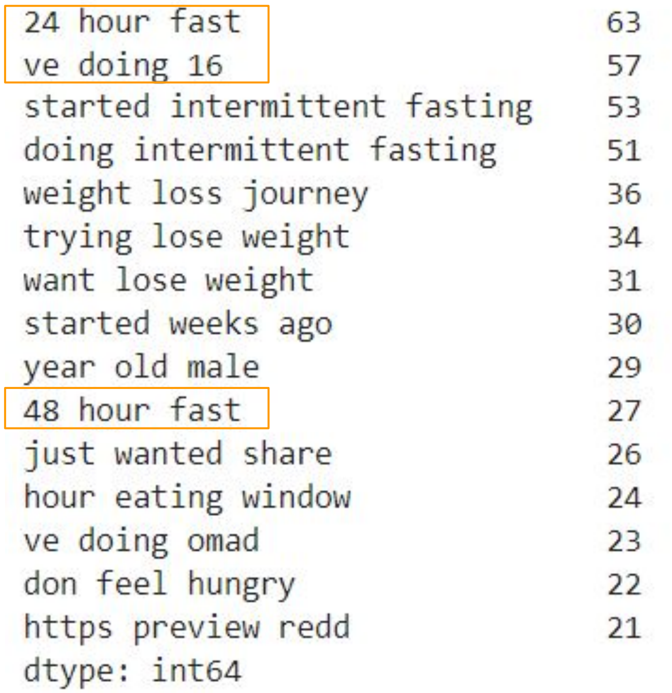
## Keto subreddit



```
weight loss       725
low carb          675
keto diet         615
doing keto        564
started keto      520
feel like         437
lose weight       421
net carbs         400
new keto          285
ve keto           283
support thread    281
blood sugar       275
don want          249
ve doing          243
don know          232
dtype: int64
```

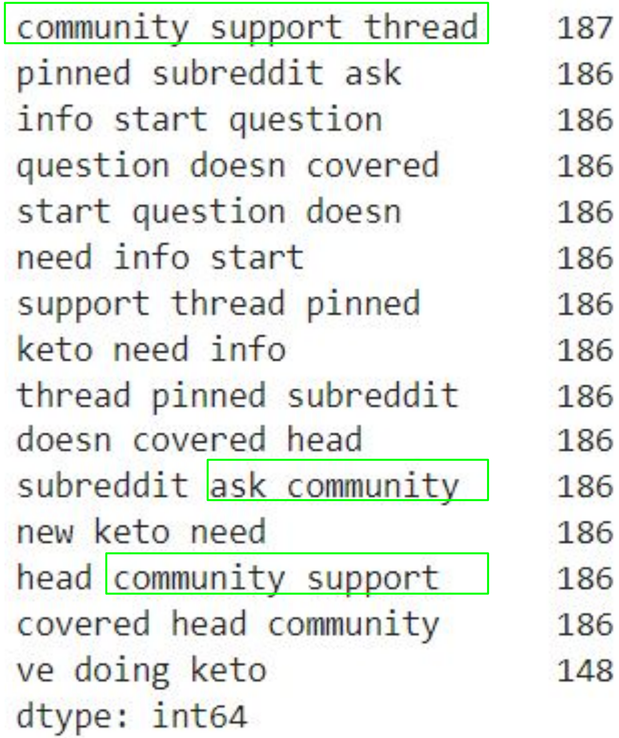
# Most common trigrams

## Intermittent fasting subreddit



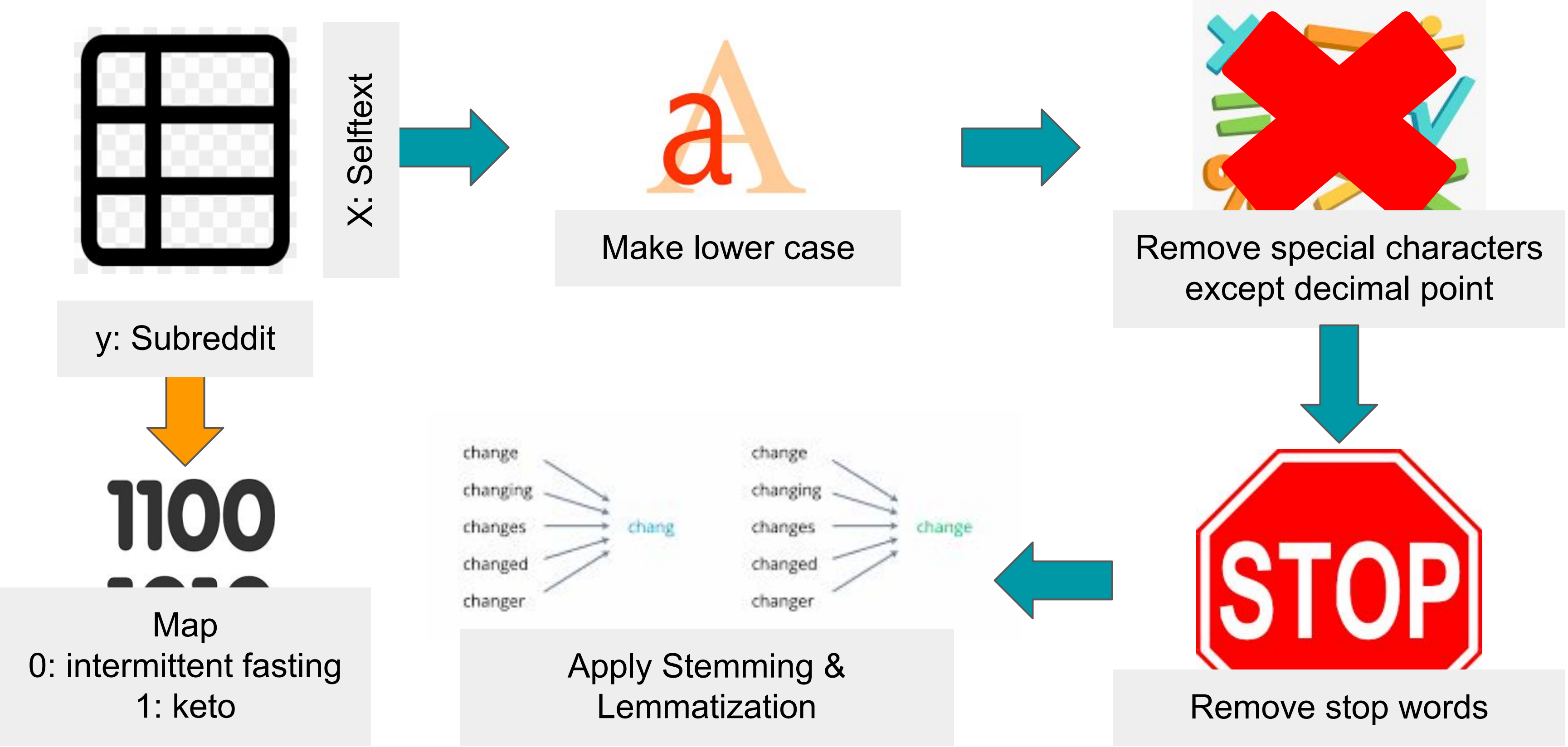
```
24 hour fast                 63
ve doing 16                  57
started intermittent fasting 53
doing intermittent fasting   51
weight loss journey          36
trying lose weight           34
want lose weight             31
started weeks ago            30
year old male                29
48 hour fast                 27
just wanted share            26
hour eating window           24
ve doing omad                23
don feel hungry              22
https preview redd           21
dtype: int64
```

## Keto subreddit



```
community support thread     187
pinned subreddit ask         186
info start question          186
question doesn covered       186
start question doesn         186
need info start              186
support thread pinned        186
keto need info               186
thread pinned subreddit      186
doesn covered head           186
subreddit ask community      186
new keto need                186
head community support       186
covered head community       186
ve doing keto                148
dtype: int64
```

# Pre-processing



X: Selftext

Make lower case

Remove special characters except decimal point

Remove stop words

Apply Stemming & Lemmatization

y: Subreddit

Map
0: intermittent fasting
1: keto

# Modeling

- Used 4 different classification models
    - Multinomial Naive Bayes
    - Random Forest
    - Ada boost classifier
    - Support Vector
- Both Countvectorizer and TF-IDF vectorizer applied for each model type

# Model evaluation metrics

**Accuracy**

Number of predictions that were correct

$$\frac{TP+TN}{TP+TN+FP+FN}$$

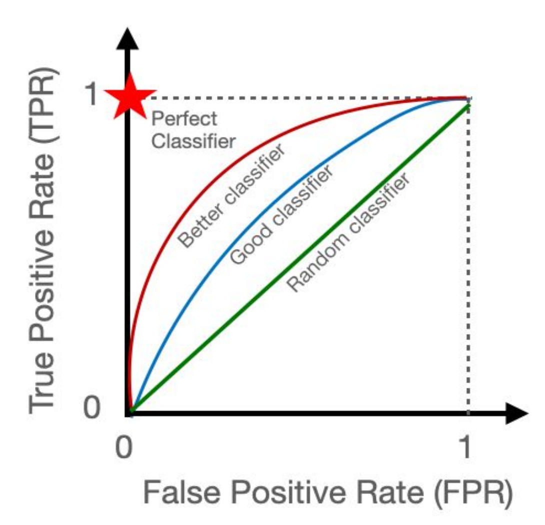
**F1 Score**

Emphasis of model was to get correct predictions. F1 score accounts for both FN and FP

$$\frac{2*(Precision*Recall)}{Precision + Recall}$$

**Receiver Operating Characteristic (ROC) Area Under Curve (AUC)**

Plots true positive rate (Sensitivity) against false positive rate (1 - Specificity)



# Model results

| Model | Accuracy (test set) | F1 score (test set) | ROC AUC |
|---|---|---|---|
| Baseline (predict majority class) | 0.561 | - | - |
| Multinomial Naive Bayes (Countvectorizer) | 0.892 | 0.902 | 0.952 |
| Multinomial Naive Bayes (TF-IDF) | 0.883 | 0.896 | 0.949 |
| Random Forest (Countvectorizer) | 0.903 | 0.918 | **0.968** |
| Random Forest (TF-IDF) | 0.903 | 0.918 | **0.968** |
| Ada Boost (Countvectorizer) | 0.892 | 0.903 | 0.961 |
| Ada Boost (TF-IDF) | 0.898 | 0.909 | 0.956 |
| Support Vector (Countvectorizer) | **0.909** | **0.919** | 0.961 |
| Support Vector (TF-IDF) | 0.905 | 0.915 | 0.965 |

# Model results

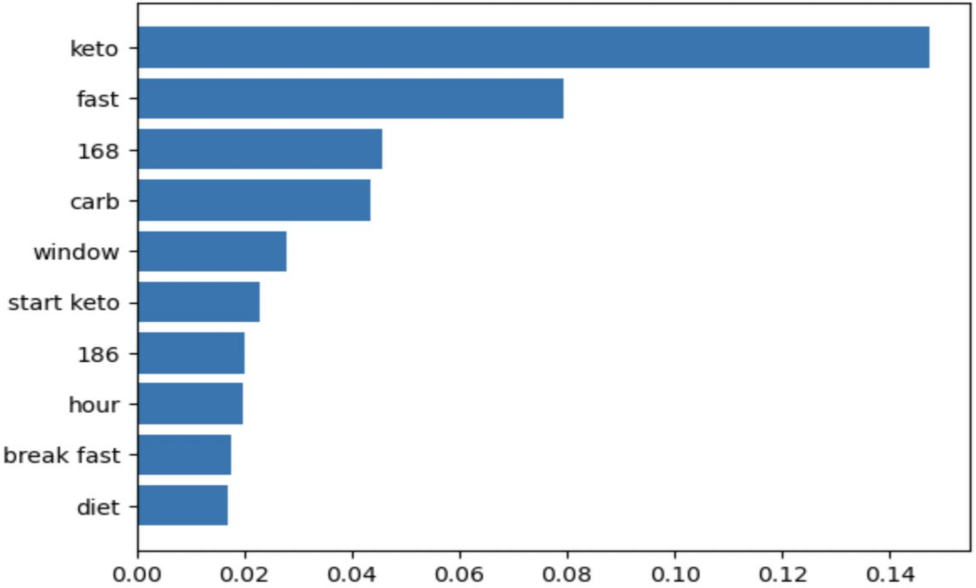
| Model | Accuracy (test set) | F1 score (test set) | ROC AUC |
|---|---|---|---|
| Baseline (predict majority class) | 0.561 | - | - |
| Multinomial Naive Bayes (Countvectorizer) | 0.892 | 0.902 | 0.952 |
| Multinomial Naive Bayes (TF-IDF) | 0.883 | 0.896 | 0.949 |
| Random Forest (Countvectorizer) | 0.903 | 0.918 | 0.968 |
| Random Forest (TF-IDF) | 0.903 | 0.918 | 0.968 |
| Support Vector (TF-IDF) | 0.905 | 0.915 | 0.965 |

**Support vector (Countvectorizer) had higher metric scores but the train set performs (accuracy: 0.959) much better than the test set (0.909), suggesting the model is overfit**

**Random Forest (TF-IDF) had similar metric score with the Countvectorizer one but the difference in score between train and test was larger**

# Best Model further analysis (Feature importance)

Based on the selected model (Count Vectorizer with Random Forest), we obtain the top 10 important (significant) features that are being utilized by the model in classifying the posts into the subreddit category (intermittent fasting and keto). We can observe the feature importance for our best selected model.



# Sentiment Analysis of selftext

We further analyzed the selftext on sentiment analysis in which we seek to classify text as having a positive, neutral or negative sentiment. We would like to understand the public sentiments in regard to intermittent fasting and keto.

**Intermittent fasting subreddit**

| 🙂 | 😐 | 😠 |
|---|---|---|
| ↓ | ↓ | ↓ |
| 67% | 5% | 28% |

**Keto subreddit**

| 🙂 | 😐 | 😠 |
|---|---|---|
| ↓ | ↓ | ↓ |
| 68% | 3% | 29% |

# Conclusion and Recommendations

With the common bigrams and trigrams which we have obtained from the exploratory data analysis stage, we compiled the following list of words for both intermittent fasting and keto in which we recommend the meal food company to consider to utilize for their advertising campaigns (i.e. hashtags for instagram, etc.)

**Words that can be used for intermittent fasting**

- weight loss
- 16:8
- 18:6
- intermittent fasting
- break fast
- diet
- lose weight
- 24 hour fast
- trying lose weight
- weight loss journey
- calories
- eat
- eating

**Words that can be used for keto**

- weight loss
- low carb
- keto diet
- doing keto
- started keto
- lose weight
- net carbs
- new keto
- eat
- eating
- keto

# Conclusions and Recommendations (continued)

Based on the sentiment analysis which we have conducted, we observed that for both the topics of intermittent fasting and keto, around 70% of the audience had positive sentiments. With this seemingly high positive response from the audience, we would recommend the meal eating company to continue to offer intermittent fasting diet plans and consider expanding to include keto meal diet plans to offer to their customers.

Both intermittent fasting and keto seems to have quite positive responses and would be something that someone can consider to adopt if they are planning to change their lifestyle to a healthier one.

What is the Keto Diet?
70% FATS
Almonds
Avocado
Coconuts
Peanuts
Olive oil
25% PROTEINS
Chicken
Tofu
Salmon
Eggs
Broccoli
5% CARBS
Bread
Corn