SUBREDDIT CLASSIFICATION WITH NATURAL LANGUAGE PROCESSING

Veronica Leong

AGENDA

- Problem Statement
- Subreddits - Nutrition & Keto
- Data Cleaning & EDA
- Modeling
- Conclusion & Recommendations
- Future Project Refinements

PROBLEM STATEMENT

How can we determine the subreddit of a new post based on the text of its Submission, using classification modeling?

SUBREDDIT CHOICES

r/ Nutrition

- nutrition science, macro/micro nutrients, health supplements, and overall diets

r/ Keto

- ketogenic diet - thoughts, experiences, and keto lifestyle advice

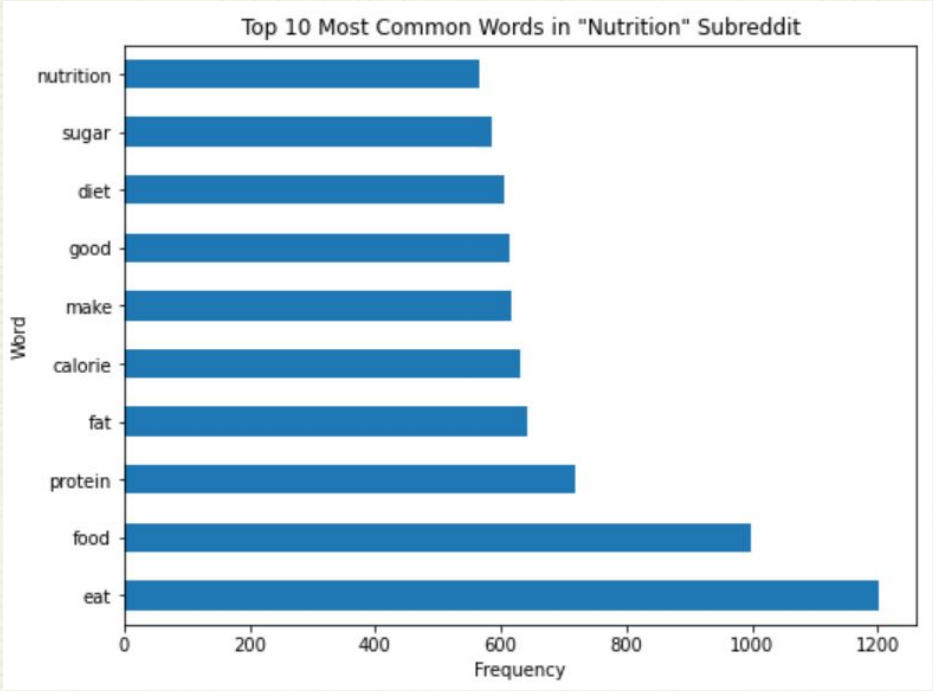
Data Sources: r/Nutrition ; r/Keto

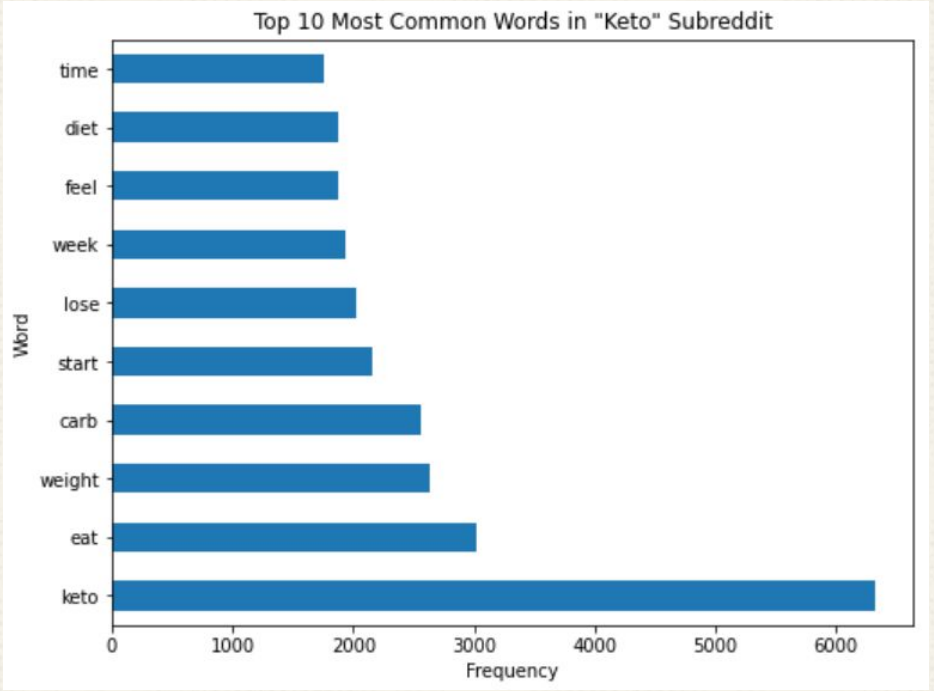
DATA CLEANING

- Collected 5,000+ rows from each subreddit
- Removed null values
- Removed duplicate ‘selftext’ records
- After cleaning → 3,000+ rows each subreddit
- Lemmatized ‘selftext’ feature

Data Sources: r/Nutrition ; r/Keto; PushShift API

SUBREDDIT EDA





Data Sources: r/Nutrition ; r/Keto

MODELING

Model	Accuracy Score
Multinomial Naive Bayes with CountVectorizer	86.05%
Multinomial Naive Bayes with TfidfVectorizer	84.84%
Logistic Regression with CountVectorizer	90.80%
Logistic Regression with TfidfVectorizer	**91.15%**
Random Forest Classifier with CountVectorizer	88.77%
Baseline Accuracy Score	*55%*

Data Sources: r/Nutrition ; r/Keto

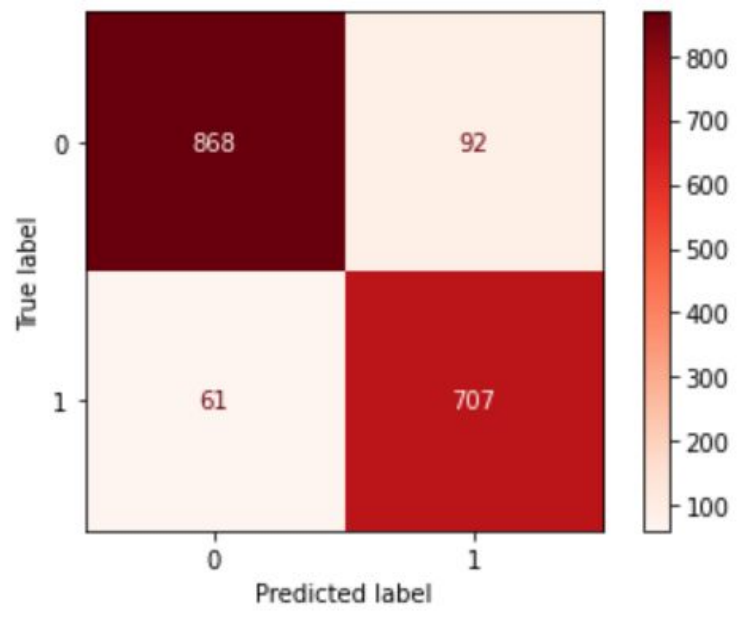
BEST MODEL

- **Logistic Regression with TifidfVectorizer**
- Additional English stop_words
  - ‘really’, ‘like’, ‘ve’, ‘don’, ‘just’, ‘day’, ‘know’, ‘does’, ‘https’
- GridSearchCV -- Best Parameters
  - Max_features = 1000
  - Ngram_range = (1, 2)

91% Accuracy

Data Sources: r/Nutrition ; r/Keto

BEST MODEL



0 = r/Keto ; 1 = r/Nutrition

- **F1-Score: 90.24%**
- **Precision: 88.49%**
- **Recall: 92.06%**

Data Sources: r/Nutrition ; r/Keto

BEST MODEL (CONT'D)

Top Distinguishable Words Per Subreddit

r/Nutrition	r/Keto
healthy	keto
nutrition	carb
vitamin	ketosis
nutrient	week
acid	start
food	start keto
oats	net
health	net carb
benefit	month
calorie	lose

Data Sources: r/Nutrition ; r/Keto

CONCLUSION & RECOMMENDATIONS

- Logistic Regression was best able to classify the ‘selftext’ to the correct Subreddit
- Most common words:
  - r/Nutrition: eat, food, protein, fat, calorie
  - r/Keto: keto, eat, weight, carb, start

Recommendations →

- Lemmatize prior to modeling
- Incorporate stop_words
- Look at different ngram_ranges

Data Sources: r/Nutrition ; r/Keto

FUTURE PROJECT REFINEMENTS

- Analyze other features (comments, titles)
- Analyze a combination of features
- Build and evaluate additional classification models
  - (KNearestNeighbors, DecisionTree)

Data Sources: r/Nutrition ; r/Keto

THANK YOU!
QUESTIONS?