DailyDietitian

Go vegan or not for nutrition is hot topics on daily Reddit Vegan and Nutrition posts. As dietitian at DailyDietitian company, we engage to provide insights by solving binary classification challenge

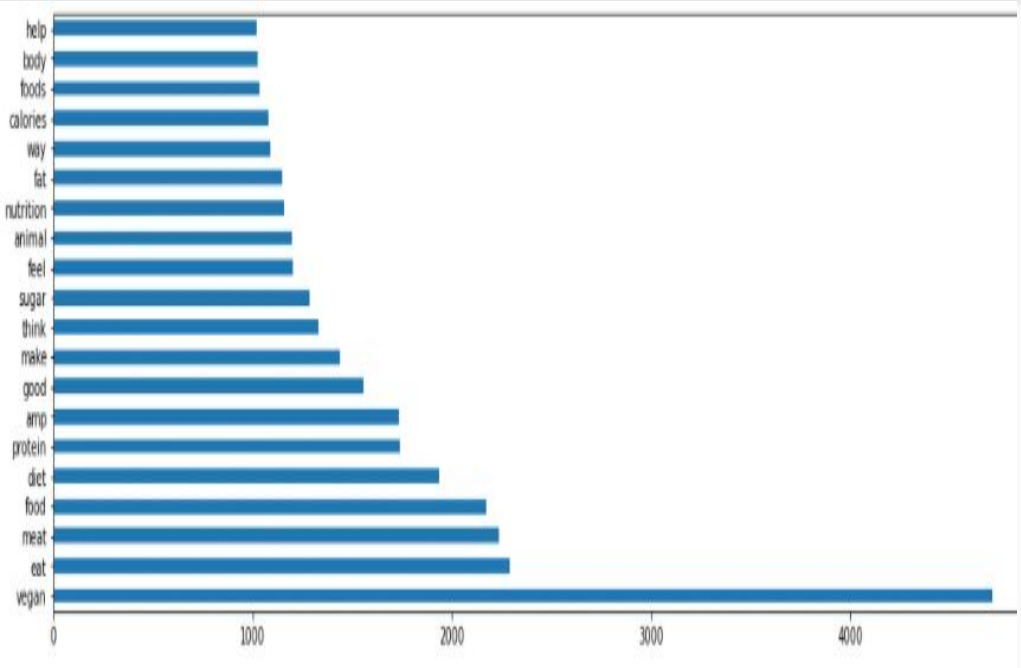
By Grace Chang

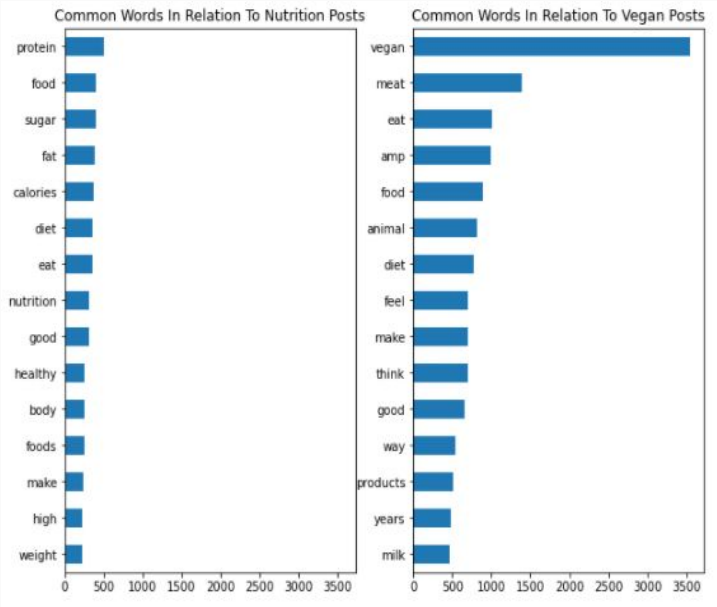
Problem Statement

How can we use predictive modeling to best predict which subreddit a post came from? Which words will most differentiate our two chosen subreddits - vegan and nutrition?

The goal of the project is to train the classifier on which subreddit a given post came from - Vegan and Nutrition to solve a binary classification problem by showcase on WebScraping | APIs | Natural Language Processing on Reddit Vegan and Nutrition posts

Identify the most repeated words
Compare them in vegan and nutrition





Classification Report and Confusion Matrix with Sensitivity and Specificity

The classifier made a total of 3058 predictions. Out of those 3058 cases, the classifier predicted vegan 1200 times and nutrition 1958 times. The accuracy is 93.79% with a sensitivity of 87.92% and specification of 97.58%

Classification Report

	precision	recall	f1-score	support
nutrition	0.93	0.98	0.95	1858
vegan	0.96	0.88	0.92	1200
accuracy			0.94	3058
macro avg	0.94	0.93	0.93	3058
weighted avg	0.94	0.94	0.94	3058

```
[[1813   45]
 [ 145 1055]]
```

Confusion Matrix

```
# Calculate Accuracy
accuracy = (tp + tn) / (tp + tn + fp + fn)
print(f'Accuracy: {round(accuracy, 4)}')
```

```
Accuracy: 0.9379
```

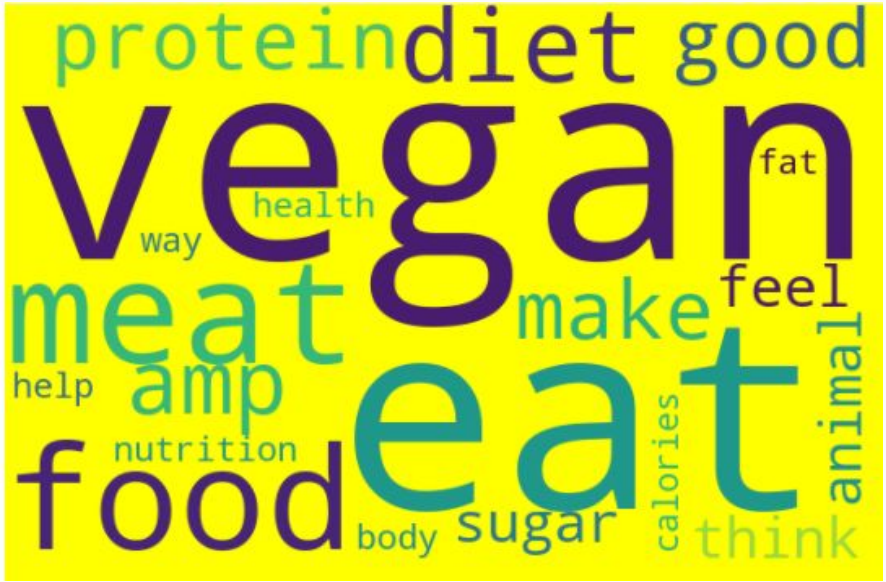
```
# Calculate sensitivity
sensitivity = tp / (tp + fn)
print(f'Sensitivity: {round(sens, 4)}')
```

```
Sensitivity: 0.8792
```

```
# Calculate specificity
specification = tn / (tn + fp)
print(f'Specificity: {round(spec, 4)}')
```

```
Specificity: 0.9758
```

Visualize top 20 common words from both vegan and nutrition in wordcloud



Summary

I have collected data, identified common words, selected models to build, evaluated the models, and compared the models with natural language processing tools. The details listed below:

The baseline: 50%

Performance metrics: accuracy and precision

The data collection method: Reddit API through pushshift.io. I collected 17266 posts in total, 12098 nutrition and 5168 vegan, from 120 days before until the day of the data collection

Vectorizers used: CountVectorizer, TfidfVectorizer to create the sparse matrix of features count/frequency respectively, to feed it to the classification model; tokenizer is included in these vectorizers

Models used/tested: Logistic Regression, Multinomial Naive Bayes, Random Forest

Modeling tools used: Pipelines and GridSearch

Evaluation methods: accuracy score, cross-validation, coefficiency, precision from classification report and confusion matrix

Conclusions / Recommendations

Conclusions:

The best score of TFDIF with Logistic Regression is: 0.9414
The best score of CountVectorizer with Logistic Regression is: 0.9352
The best score of TFDIF with naive multinomial Bayes is: 0.9187
The best score of random foresets is: 0.919

Recommendations:

Based on the conclusions, the winning model was the TFDIF with Logistic Regression that has a 94% accuracy score. We recommend to go for it.

Data Sources

Vegan Data:
Vegan Data collected with 120 days period from Reddit to satisfy the requirements:
base_v_url = 'https://api.pushshift.io/reddit/search/submission/?subreddit=vegan&size=200&before={}d'
vegan_urls = [base_v_url.format(i) for i in range(120,-1,-1)]

Nutrition Data:
Nutrition Data collected with 120 days period from Reddit to satisfy the requirements:
base_n_url = 'https://api.pushshift.io/reddit/search/submission/?subreddit=nutrition&size=200&before={}d'
nutrit_urls = [base_n_url.format(i) for i in range(120,-1,-1)]