Project 3:
Web APIs & Classification

Building a classifier to identify subreddit postings

Rachel Lim - DSI 16

Agenda

- Business Problem
- Methodology
- EDA
- Model & Performance
- Limitations
- Conclusion

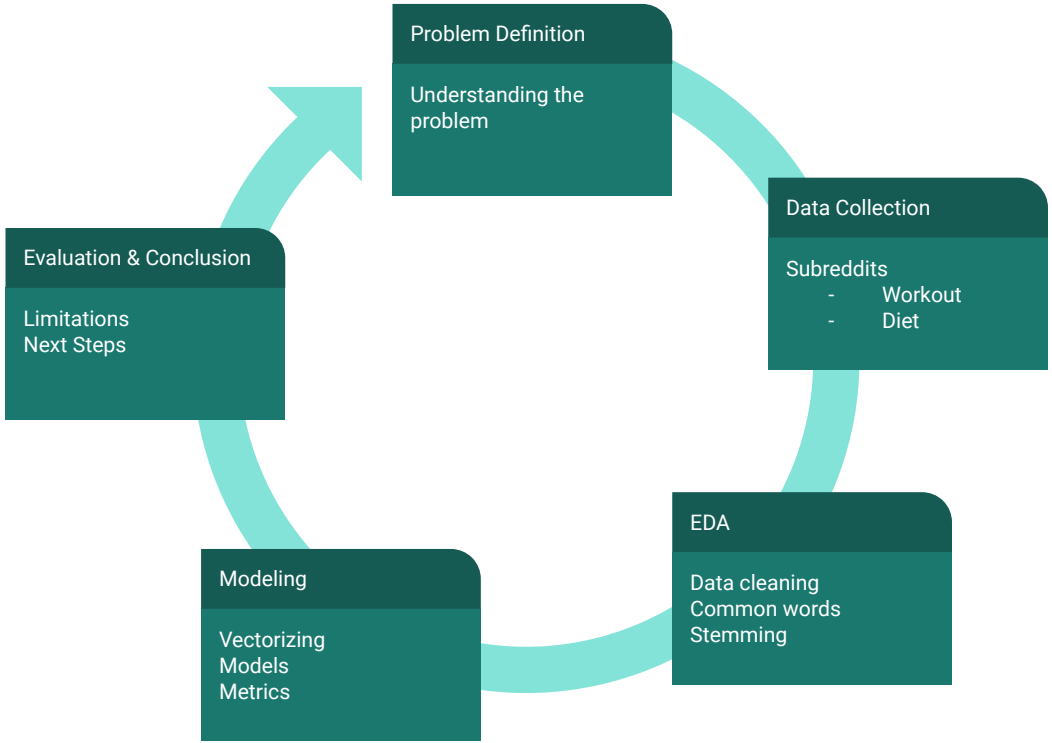
Business Problem

All Wellness is a popular online platform aiming to help people improve their overall well being via different channels. One of which is having certified fitness and nutrition coaches giving advices to the platform members on the fitness and dietary queries they have in their workout routine or nutrition and diet.

One issue All Wellness faced very often is that when members wants to get advice, they typically do not label the nature of their question or label wrongly, and it takes time and effort to manually classify the queries.

This project aims to use NLP to classify if the query is fitness related, or diet related, tag them and channel the query to their panel of certified fitness coach or nutritionist to answer.

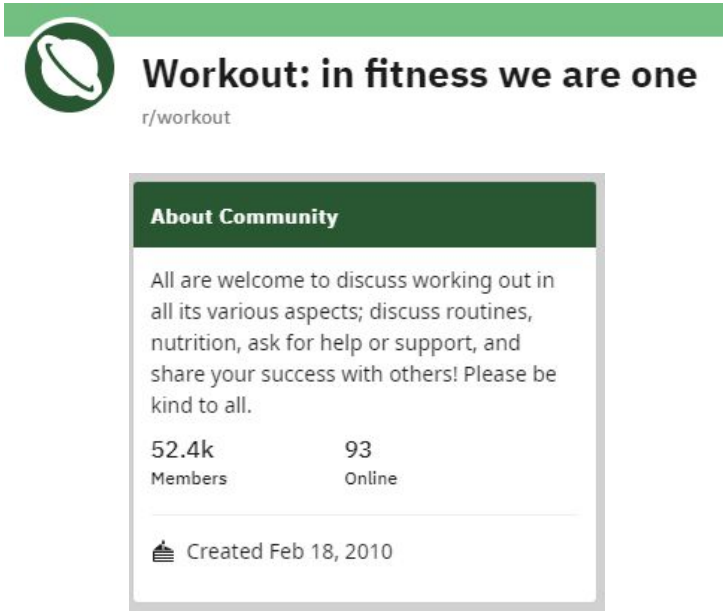
Methodology



Methodology

- Data Source

For the data, we will use Reddit's API to scrape 2 subreddits for posts (which includes text and title) and use NLP to train a classifier to identify the type of query.



EDA

- 998 text records for /r/workout in 'workout.csv'
- 998 text records for /r/diet in 'diet.csv'
- After removing duplicates or empty posts
- we have ~917 vs 913 posts for each post (balance data)

	subreddit	post
0	workout	Do you need to Gain Weight, Lose Weight, or Ma...
1	workout	Beginner's Guide to Working Out As a personal...
2	workout	By next Saturday, 1 upvote = 5 pushups, will p...
3	workout	Morning workout in my adidas. What are yall do...
4	workout	Day 25 of Zac Efron's 12 Week Baywatch Programme
...	...	...
1828	diet	Am I
1829	diet	50 ca
1830	diet	BEST WHEY PRO
1831	diet	Keep g
1832	diet	Why is

1833 rows × 2 columns

		subreddit	diet	workout
post	count		917	913
	unique		917	913
	top		Turmeric and how to use it Does putting a pinc...	Dont feel the burn Ive been working out at the...
	freq		1	1
wordslist	count		917	913
	unique		913	913
	top		[keto]	[squeez, chest, f, small, chest, strength, tra...
	freq		3	1

- Then we proceed to do some cleaning.
  a. Removing HTML text
  b. Removing Non Letters
  c. Change to lower case
  d. Remove Stop words (common words, words related to topic)
  e. Stemming
- Labelling - Diet - 1; workout - 0

```
def cleanwords(data, do_all=True):
    """
    Function will remove all irrevant characters(or digits), HTML tags, stopwords),
    then perform stemming to get the root word.
    Takes in arguement:
    data- str the raw text to be converted

    Return the list of cleaned and stemmed words
    """
    # 1. Remove HTML.
    review_text = BeautifulSoup(data).get_text()

    # 2. Remove non-Letters.
    letters_only = re.sub("[^a-zA-Z]", " ", review_text)

    # 3. Convert to lower case, split into individual words.
    words = letters_only.lower().split()

    # 4.remove stop words, including title of the subreddits
    stops = set(stopwords.words('english'))
    stops.update(['diet', 'diets', 'dieting', 'dieter', 'dieters', 'dietary', 'dieted', 'workout', 'worl
    meaningful_words = [w for w in words if w not in stops]

    #5 Stemming
    p_stemmer = PorterStemmer()
    meaningful_words_stemmed = [p_stemmer.stem(w) for w in meaningful_words]

    #6 return the list of stemmed and cleaned words
    return(meaningful_words_stemmed)
```

EDA

- Word cloud

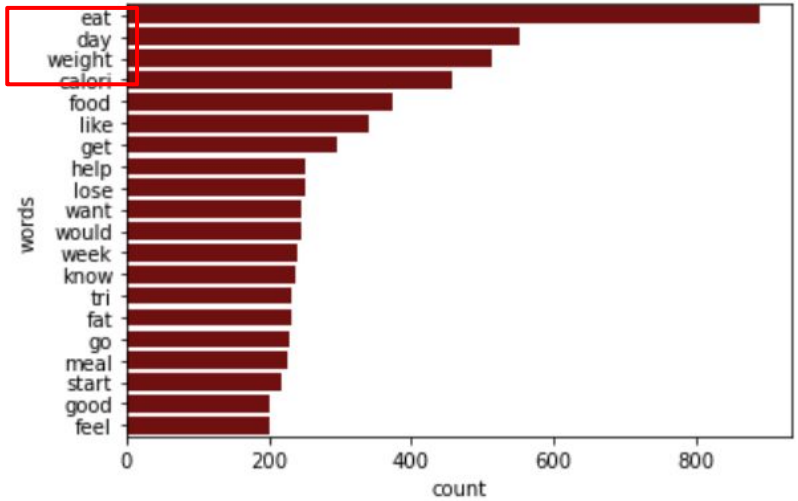
Diet Set

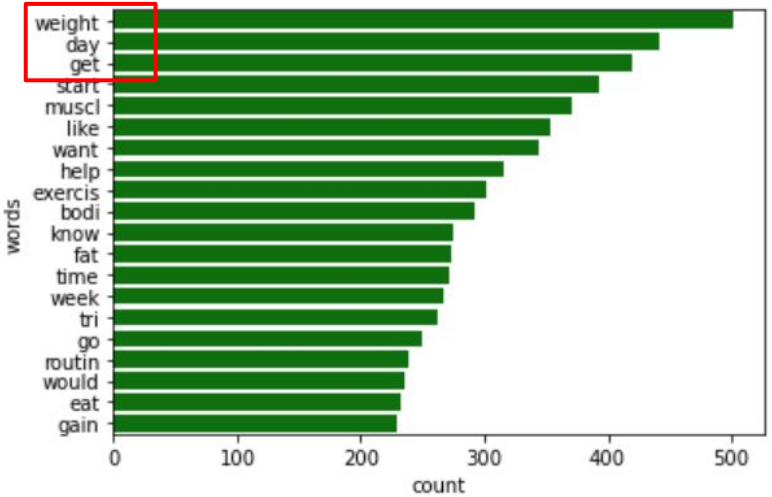
Workout Set

EDA

Common Words (to be removed)

	words	count_x	count_y	_merge
0	eat	890.000	232.000	both
1	day	552.000	442.000	both
2	weight	515.000	502.000	both
5	like	341.000	353.000	both
6	get	295.000	420.000	both
7	help	252.000	316.000	both
9	want	245.000	344.000	both
10	would	245.000	236.000	both
11	week	241.000	267.000	both
12	know	238.000	275.000	both
13	tri	233.000	262.000	both
14	fat	233.000	273.000	both
15	go	229.000	249.000	both
17	start	217.000	393.000	both





Methodology

- Models & Performance

Models:

- Logistic Regression with CountVectorizer, TF-IDF
- Naive Bayes with CountVectorizer, TF-IDF
- KNearestNeighbours with (KNN) CountVectorizer, TF-IDF
- Support Vector Machines (SVM) with CountVectorizer, TF-IDF

Metrics:

- Accuracy
  - measure of how accurate our model is:
- Matthews Correlation Coefficient (MCC)
  - The higher the correlation between true and predicted values (for both classes), the better the prediction.

The model with highest Accuracy and MCC score for on the validation data set will be deployed

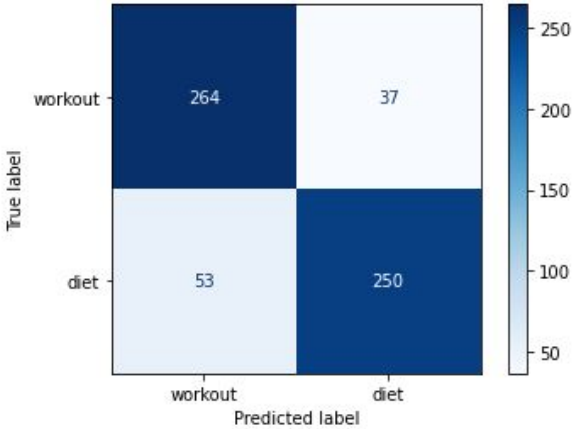
Models and Performance

Model with HyperParameters Tuning	Best Score	Accuracy on Train Data	Accuracy on Validn Data	MCC Score	Parameters
Logistic Regression with TF-IDF	**0.883**	**0.954**	**0.851**	**0.703**	{'lr__C': 1.0, 'lr__penalty': 'l2', 'lr__solver': 'liblinear', 'tvec__max_df': 0.3, 'tvec__max_features': 3000, 'tvec__ngram_range': (1, 2)}
Naive Bayes with TF-IDF	0.885	0.952	0.844	0.691	{'tvec__max_df': 0.3, 'tvec__max_features': 3000, 'tvec__ngram_range': (1, 2)}
SVM with TF-IDF	0.877	0.998	0.844	0.691	{'svm__C': 3.0, 'svm__degree': 3, 'svm__kernel': 'rbf', 'tvec__max_df': 0.3, 'tvec__max_features': 5000, 'tvec__ngram_range': (1, 2)}
KNN with TF-IDF	0.813	0.843	0.800	0.616	{'knn__n_neighbors': 13, 'tvec__max_df': 0.3, 'tvec__max_features': 5000, 'tvec__ngram_range': (1, 2)}
Logistic Regression with CountVectorizer	0.694	0.850	0.694	0.393	{'cvec__max_df': 7, 'cvec__max_features': 5000, 'lr__C': 0.01, 'lr__penalty': 'l2', 'lr__solver': 'liblinear'}
Naive Bayes with CountVectorizer	0.706	0.864	0.690	0.383	{'cvec__max_df': 7, 'cvec__max_features': 5000}
SVM with CountVectorizer	0.675	0.906	0.654	0.316	{'cvec__max_df': 7, 'cvec__max_features': 5000, 'svm__C': 3.0, 'svm__degree': 3, 'svm__kernel': 'rbf'}
KNN with CountVectorizer	0.559	0.657	0.579	0.173	{'cvec__max_df': 7, 'cvec__max_features': 1000, 'knn__n_neighbors': 7}
Baseline (Majority Class)	0.502				Majority Class (diet class)

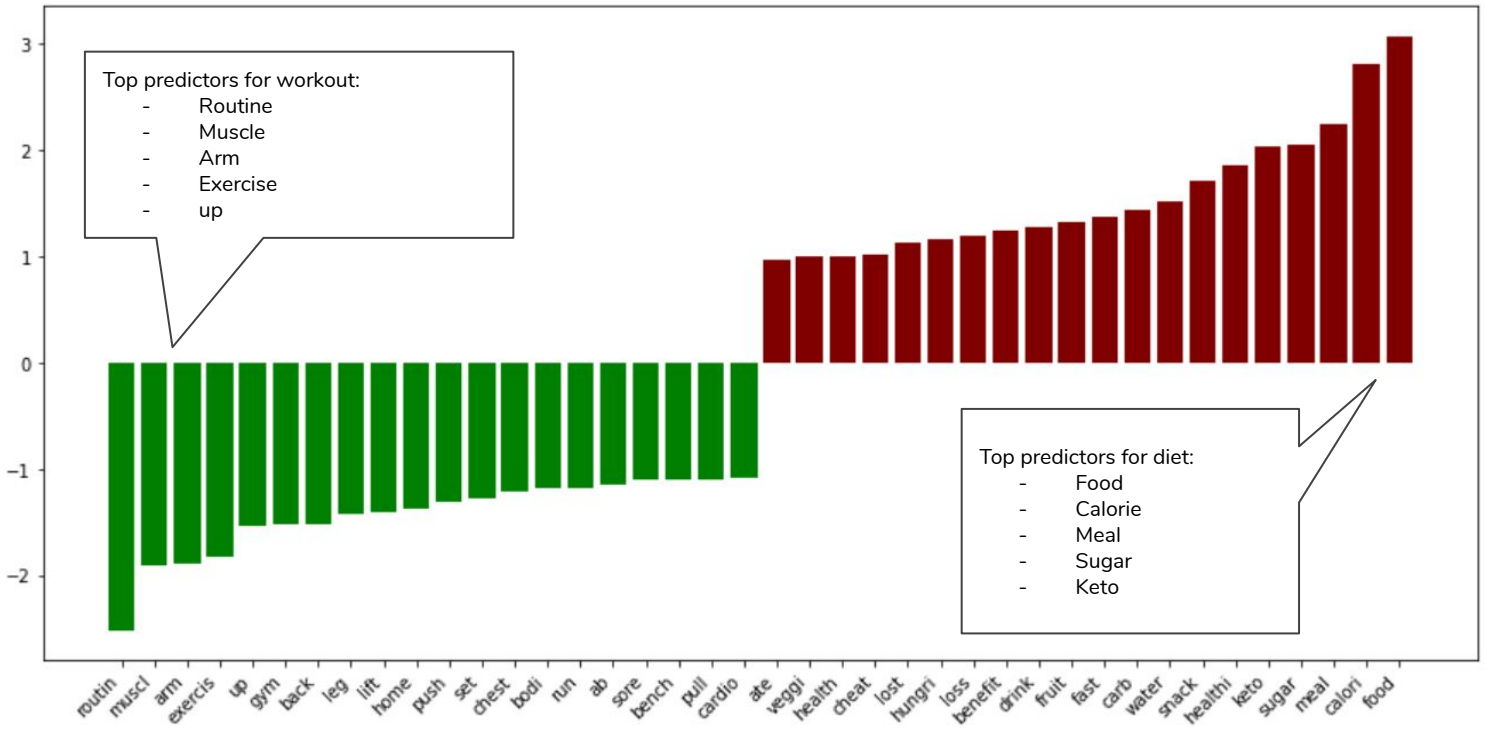
Logistic Regression

```
======== LogisticRegression +  TF-IDF with Hyperparameters Tuning =================

best parameters       :
{'lr__C': 1.0, 'lr__penalty': 'l2', 'lr__solver': 'liblinear', 'tvec__max_df': 0.3, 't
vec__max_features': 3000, 'tvec__ngram_range': (1, 2)}

best score                : 0.883
Training Accuracy         : 0.954


Validation Data Accuracy  : 0.851
Misclassification Rate    : 0.149


Matthews Correlation Coeff: 0.703
```

	precision	recall	f1-score	support
0	0.83	0.88	0.85	301
1	0.87	0.83	0.85	303
accuracy			0.85	604
macro avg	0.85	0.85	0.85	604
weighted avg	0.85	0.85	0.85	604



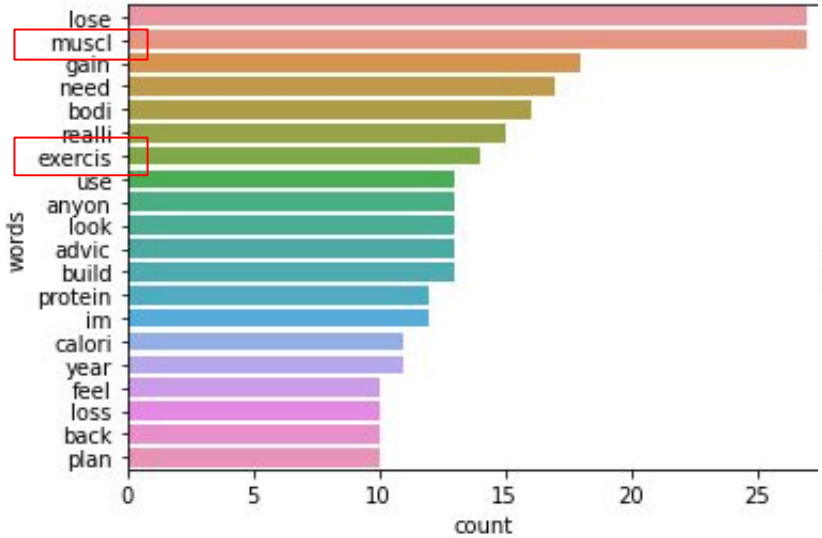
Top Predictors



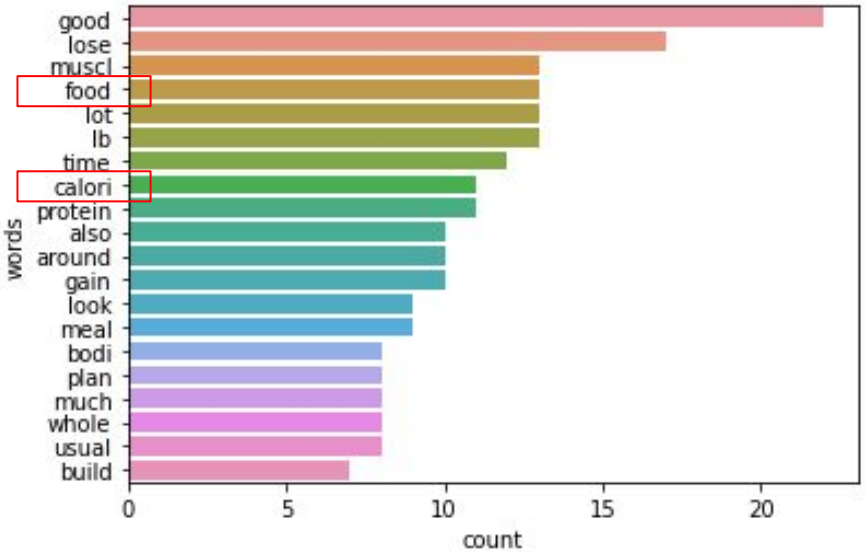
Limitations

Misclassifications

Diet posts misclassified as workout



Workout posts misclassified as diet



Limitations

Misclassifications occur when:

- Queries are too short
- Have generalized questions
- Mixed description (on working out) with dieting queries or vice versa

"Eating enough during a bulk Hey guys. I'm **doing my first bulk** this winter and I'm trying to ease myself into the eating for the first week or two. With this, I'm struggling to eat enough calories to put myself in a surplus.Based on what my fitbit says (which I take with a grain of salt) I'm burning approx. 3200 cal/day. So I'm trying work my way up to between 3300-3400 cal/day. Any suggestions on getting the calories in? **I'm trying to do a lean bulk** so I'm sticking to as much whole foods as I can."

Misclassified as Diet
Mixed queries/description

Muscle Building Carbs

Misclassified as workout
Too short post

"**Can't figure out the right way for me to eat** I'm 17, 5'3 and 115lbs. But I'm "skinny fat", so I'm trying to lose the fat and build a little bit of muscle. I'm vegan, and starting a couple days ago, I'm trying to consume 50g+ of protein. I don't like eating to be honest, but I'm doing my best so that I can hit my protein intake. My **eating habits** have been really clean for the past month and a half (or so). I exercise almost everyday as well. But it feels like no matter what, I can't lose the extra fat. Building muscle underneath isn't as hard, but fat loss feels nearly impossible at this point. Its all definitely taking a toll on my mental health, so any advice is appreciated."

Misclassified as workout
Mixed queries/description

I've been **working out at the gym (weight training)** or at home 3 times a week, have been doing cardio (I do jazz dance) 2 times a week and maintained **a low carb diet** with a max. of 1.500 calories a day for some months now. I burn around 2.500 calories a day, Instead of losing weight, I gained 3 kilos (I went from 80 to 83). I don't know what to do anymore. I honestly don't want to **starve myself** to lose weight. I don't know why I can't lose weight, according to my doctor, I am 100% healthy, so that's not the issue. Does anyone have any ideas what to do?'

Misclassified as Diet
Generalized Qn

Conclusion

Classification Model:

- Logistic Regression with TFIDF give us the highest accuracy (0.85) and MCC (0.703) score
- This will be a good start to see how much time we can cut down from tagging the queries, and how fast we can channel the queries to the relevant parties.

Next Steps:

- More data Collection (scrapping from comments section) and pre-processing
- Fine tune accuracy using Ensembling of different models
- Sentiment Analysis/Topic Modelling (to be able to predict more accurately)

Final Goal:

- Chatbot creation

Thank you