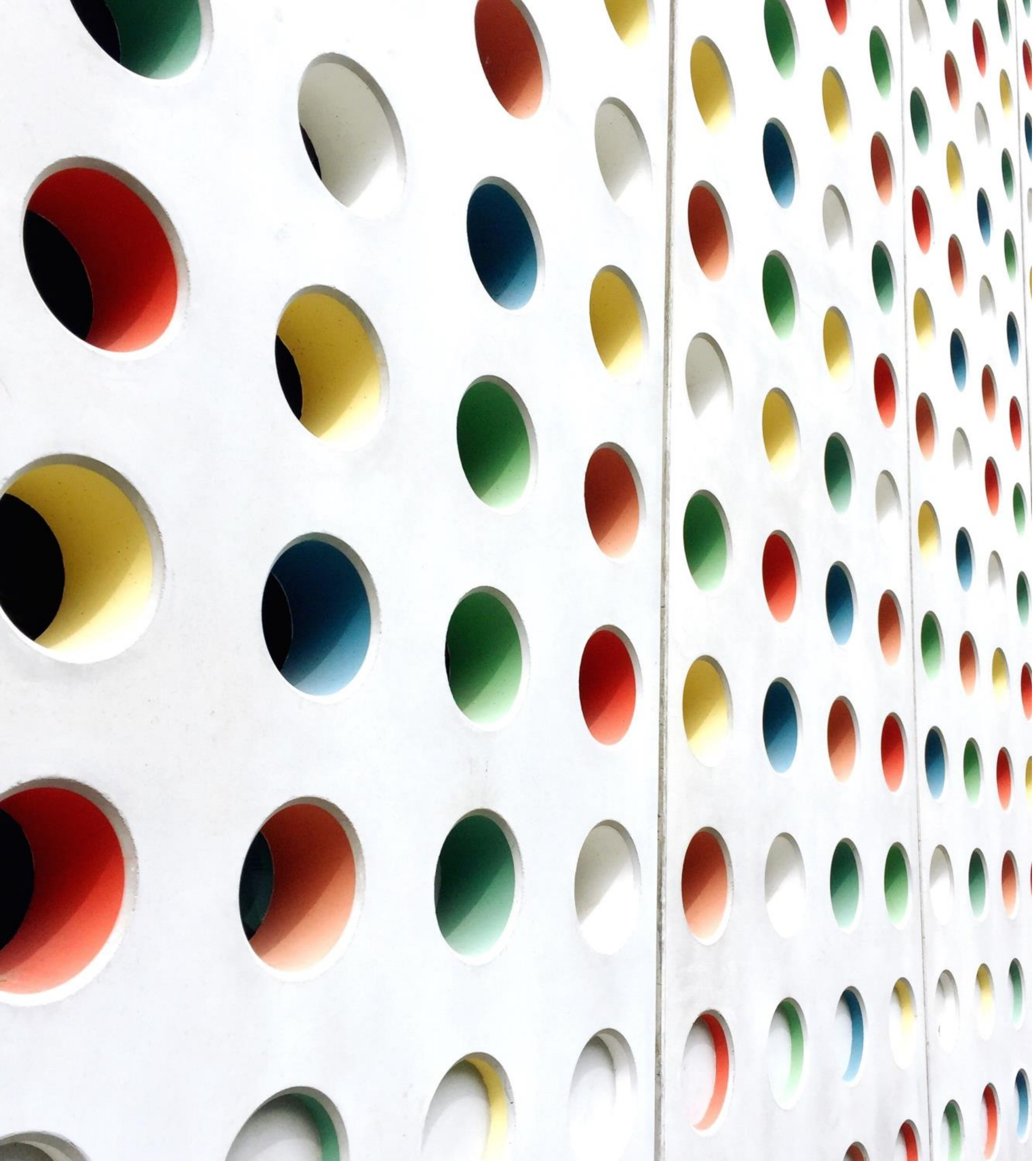


Subreddit Classification Through NLP

r/AskScience vs r/AskSocialScience

Rifqi Alkhatib
GA DSI 20 Project 3

Contents

- INTRODUCTION & PROBLEM STATEMENT
- SCRAPING
- DATA CLEANING
- TEXT PREPROCESSING
- EDA
- CREATING CLASSIFICATION MODEL
- MODEL COMPARISONS & METRICS
- LIMITATIONS
- CONCLUSIONS & RECOMMENDATIONS

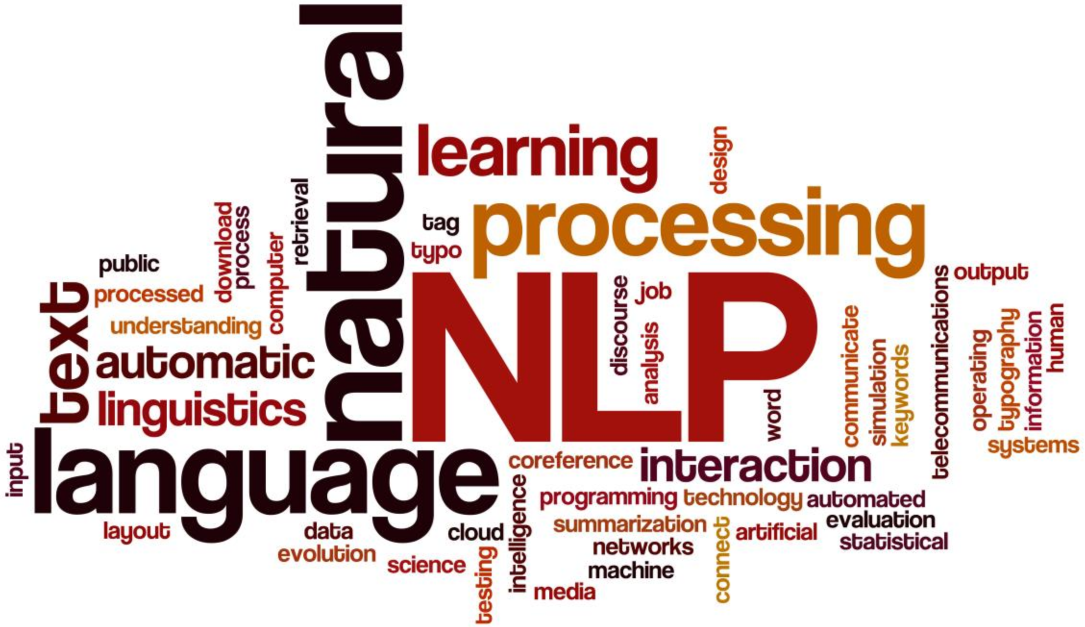
Introduction

NLP

Natural Language Processing

Help Computers Understand Natural Language

Wide Range of Applications



Introduction & Problem Statement

r/AskScience

reddit

r/AskSocialScience

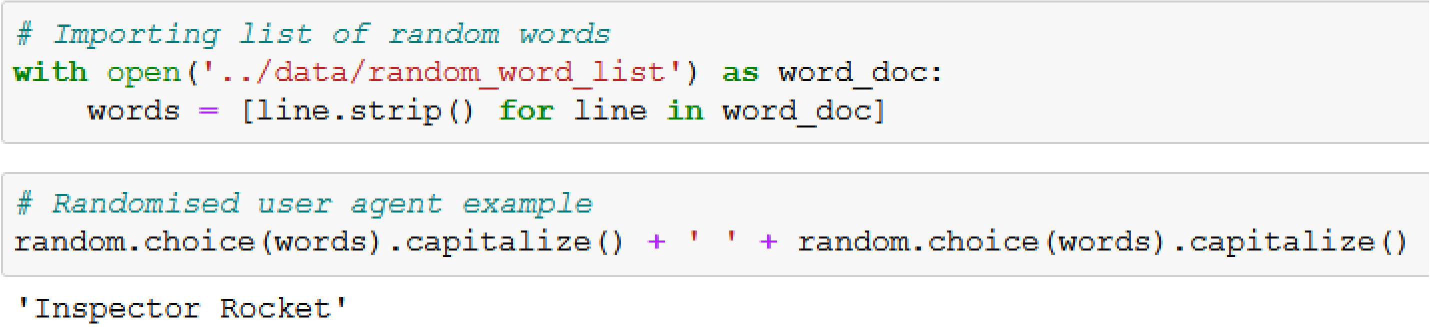
“How can we best develop a classification model using NLP to classify posts belonging to two different subreddits?”

Scraping Reddit

Reddit API

- 25 posts per request
- Max 1000 posts

Randomise User Agent



```
# Importing list of random words
with open('../data/random_word_list') as word_doc:
    words = [line.strip() for line in word_doc]
```

```
# Randomised user agent example
random.choice(words).capitalize() + ' ' + random.choice(words).capitalize()
```

```
'Inspector Rocket'
```

‘Sleep’ between requests to look more natural

Data Cleaning

Duplicates

- Overlap between requests
- Reddit API reset

Null Imputation

- Images, Videos

Moderator Posts

- Facilitating AMAs, events (KEEP)
- Weekly automated posts (REMOVE)

Text Preprocessing

Cleaning text

- html, Reddit usernames, non-alphanumeric
- Regex

Lower case

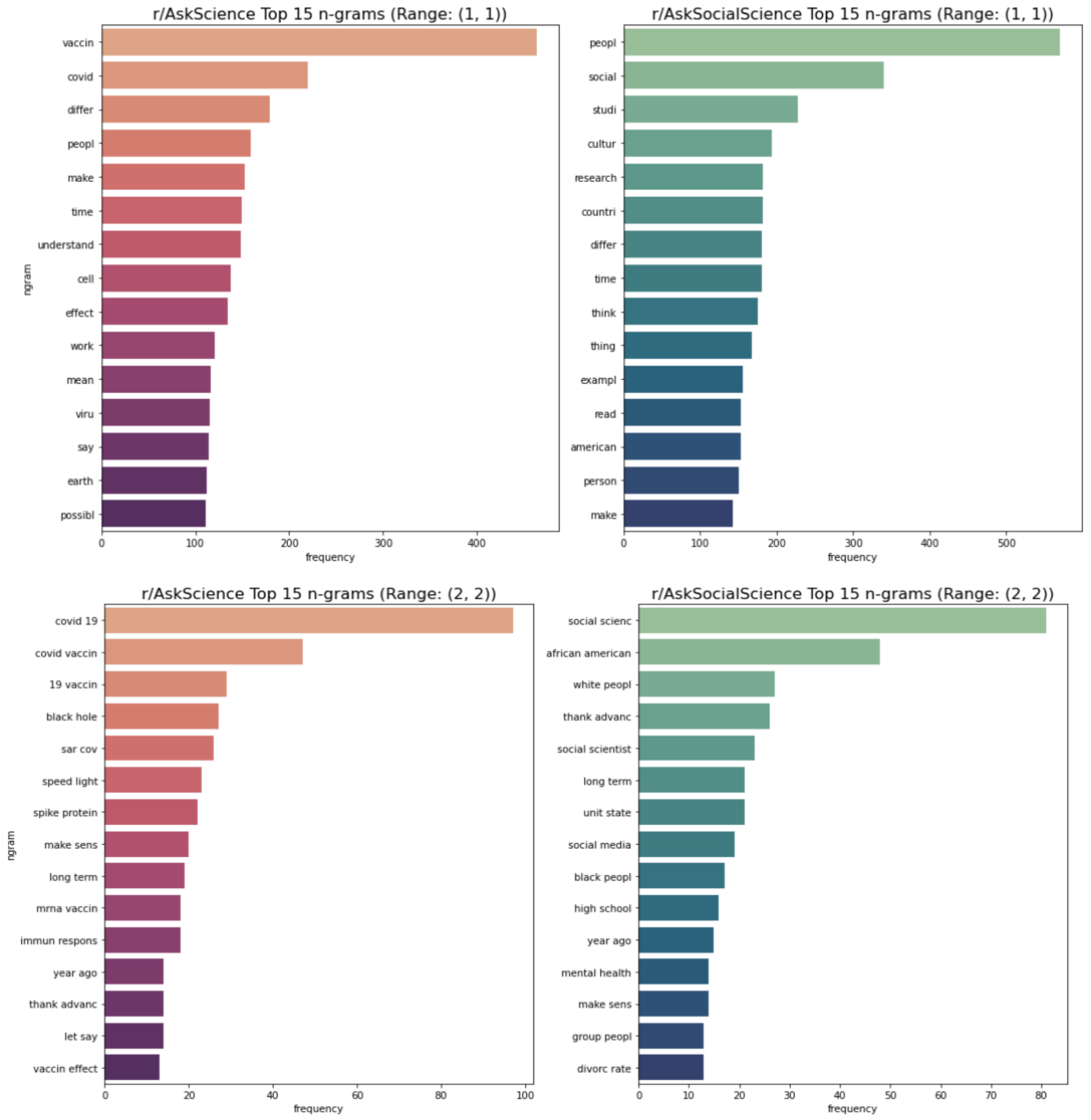
Stop Words

- 'Dead Giveaways'
- Too frequent – not meaningful

Stemming

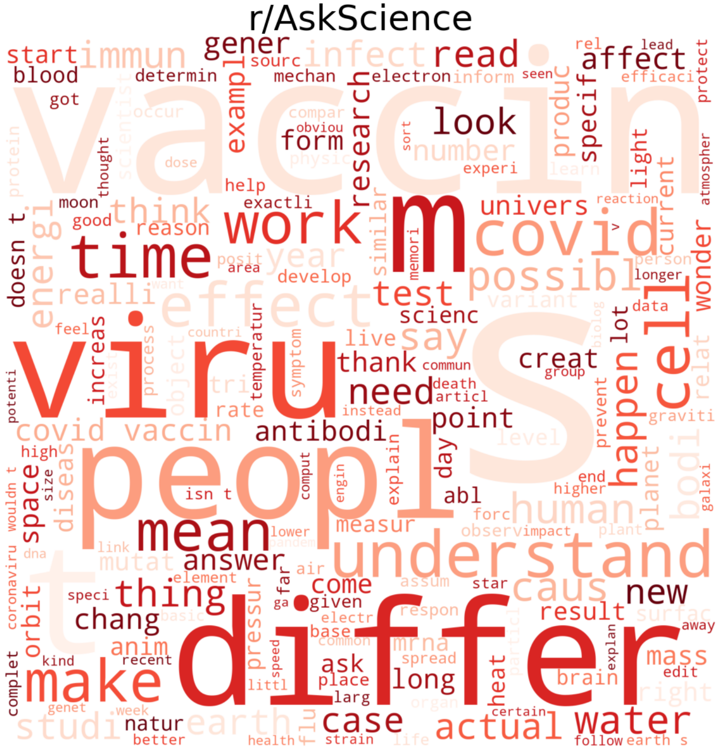
Exploratory Data Analysis

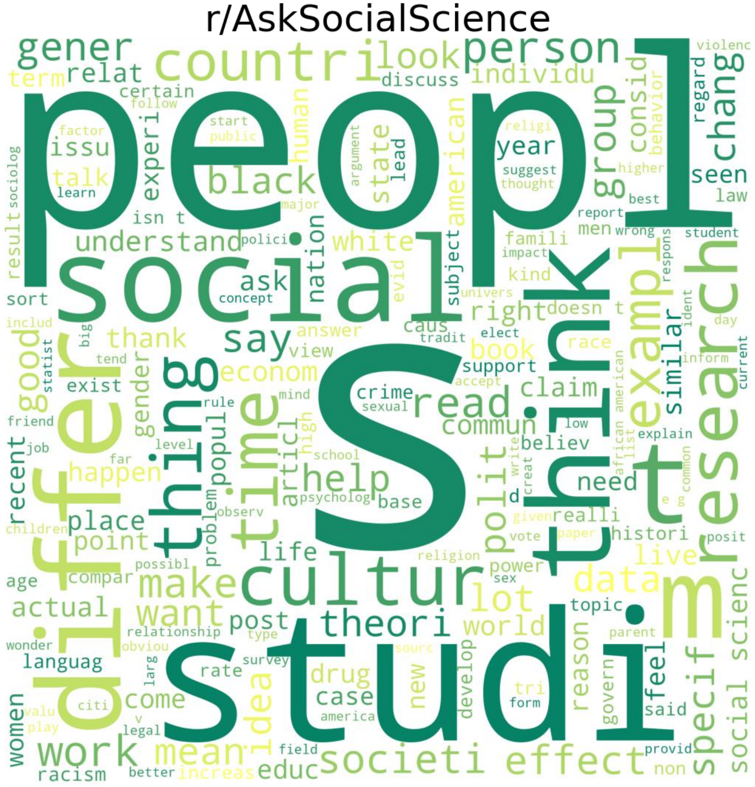
FREQUENCY OF N-GRAMS



Exploratory Data Analysis

EXPLORATORY WORD CLOUDS


r/AskScience

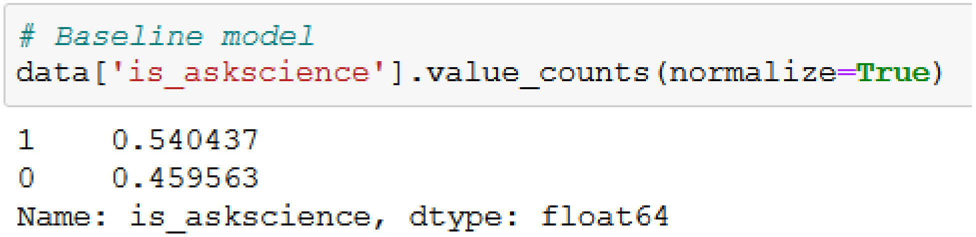

r/AskSocialScience

Creating Classification Model

1. Baseline Model
2. Train Test Split
3. Pipeline: Vectorizer & Classifier
4. Tune hyperparameters (GridSearchCV)
5. Fit model to training data
6. Evaluate
7. Compare top 2 models

Baseline Model

Actual distribution as Accuracy Score

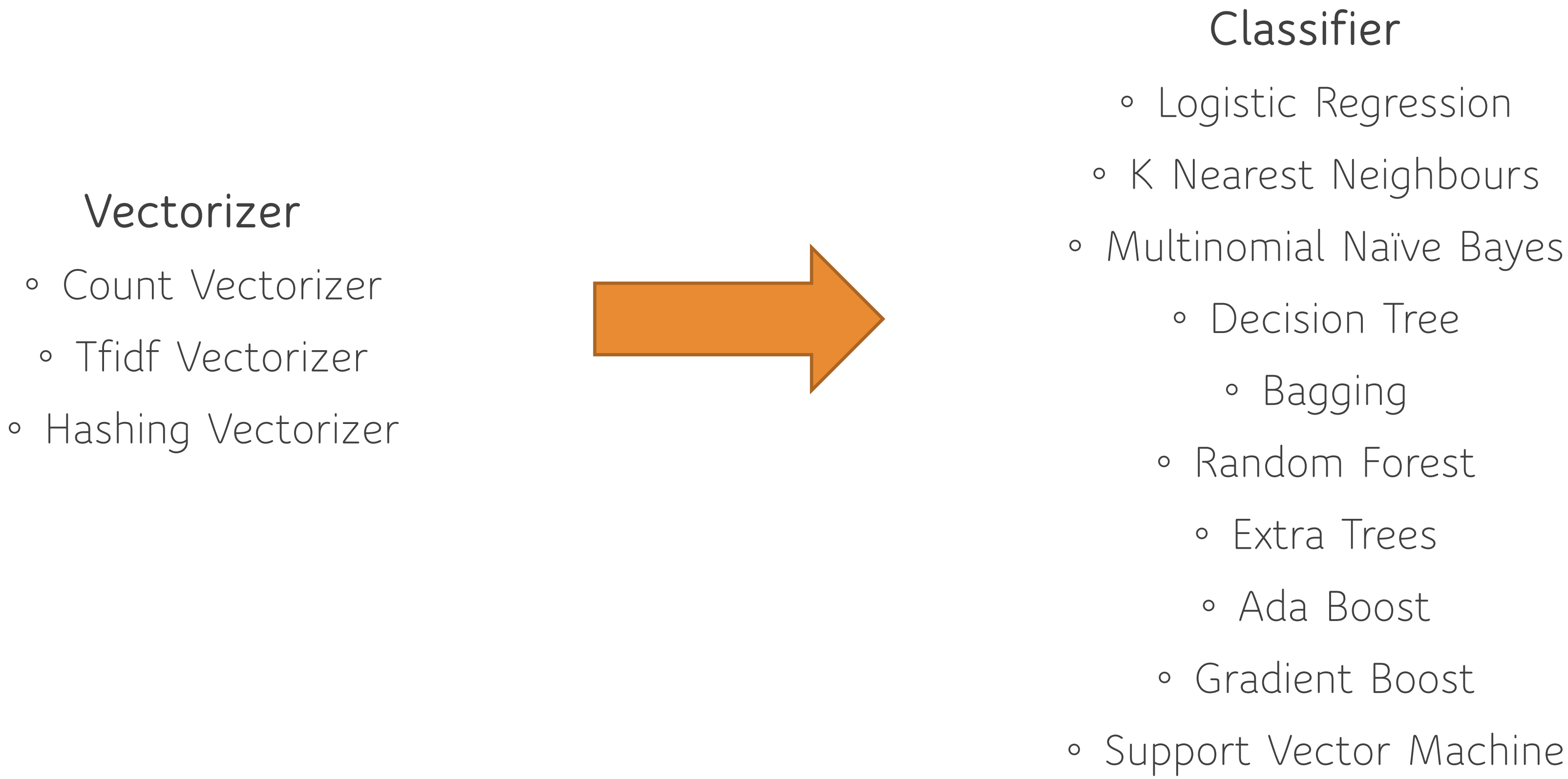


```
# Baseline model
data['is_askscience'].value_counts(normalize=True)
```

```
1    0.540437
0    0.459563
Name: is_askscience, dtype: float64
```

MUST BEAT!!!

Pipeline



Vectorizer

- Count Vectorizer
- Tfidf Vectorizer
- Hashing Vectorizer

Classifier

- Logistic Regression
- K Nearest Neighbours
- Multinomial Naïve Bayes
- Decision Tree
- Bagging
- Random Forest
- Extra Trees
- Ada Boost
- Gradient Boost
- Support Vector Machine

Tuning Hyperparameters

Conservative with parameter grid

- Initial assessment
- Computation-heavy
    - Boosting & Decision Trees

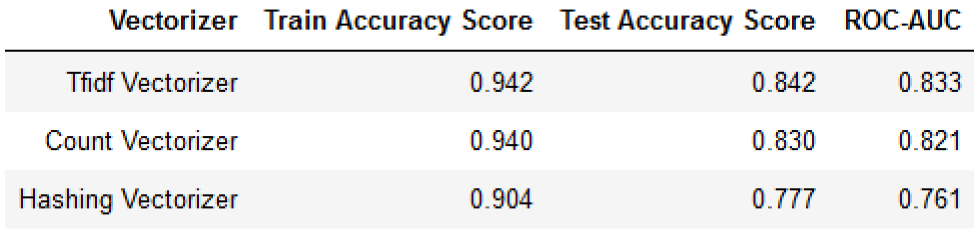
Fit to training data

Compile results for comparison

- Train & Test Accuracy
- Precision & Recall
- F1 Score
- ROC-AUC Score

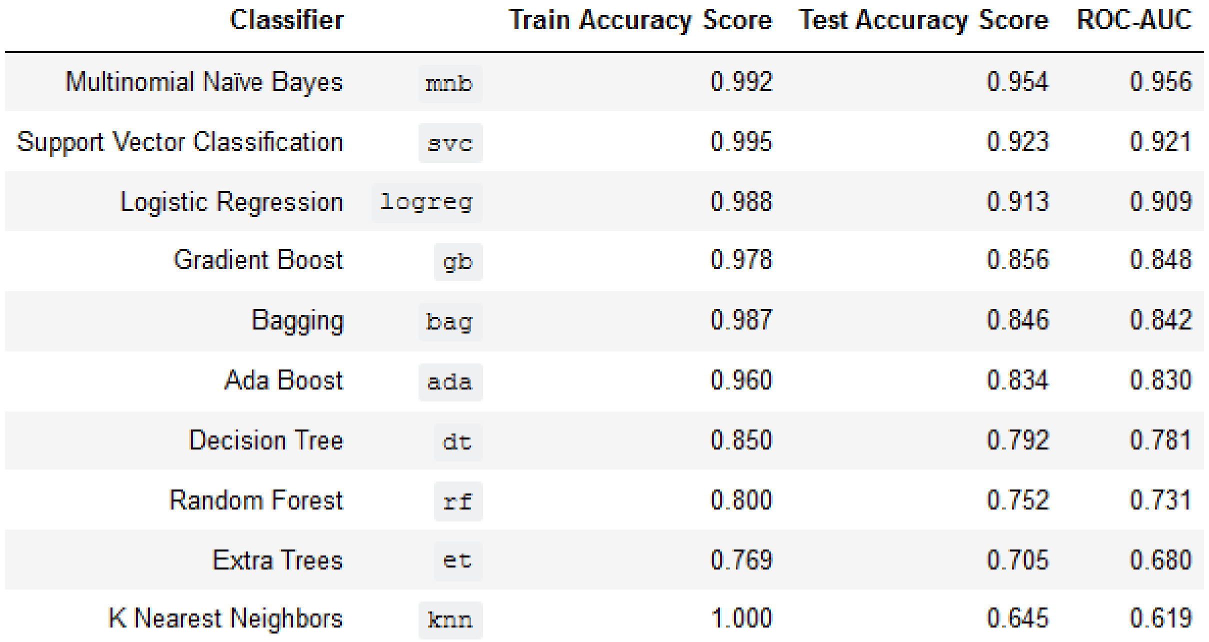
Evaluation

Vectorizer Average Scores

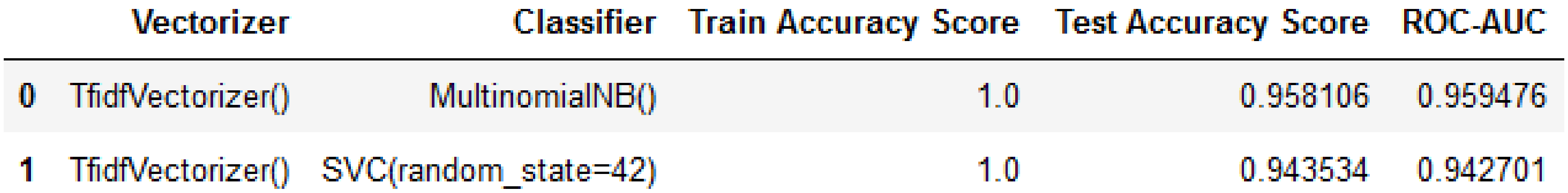


Vectorizer	Train Accuracy Score	Test Accuracy Score	ROC-AUC
Tfidf Vectorizer	0.942	0.842	0.833
Count Vectorizer	0.940	0.830	0.821
Hashing Vectorizer	0.904	0.777	0.761

Classifier Average Scores



Classifier		Train Accuracy Score	Test Accuracy Score	ROC-AUC
Multinomial Naïve Bayes	`mnb`	0.992	0.954	0.956
Support Vector Classification	`svc`	0.995	0.923	0.921
Logistic Regression	`logreg`	0.988	0.913	0.909
Gradient Boost	`gb`	0.978	0.856	0.848
Bagging	`bag`	0.987	0.846	0.842
Ada Boost	`ada`	0.960	0.834	0.830
Decision Tree	`dt`	0.850	0.792	0.781
Random Forest	`rf`	0.800	0.752	0.731
Extra Trees	`et`	0.769	0.705	0.680
K Nearest Neighbors	`knn`	1.000	0.645	0.619

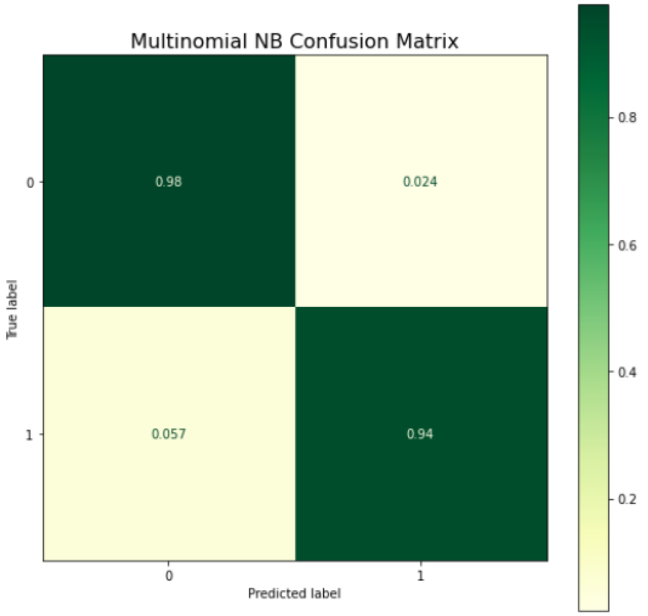


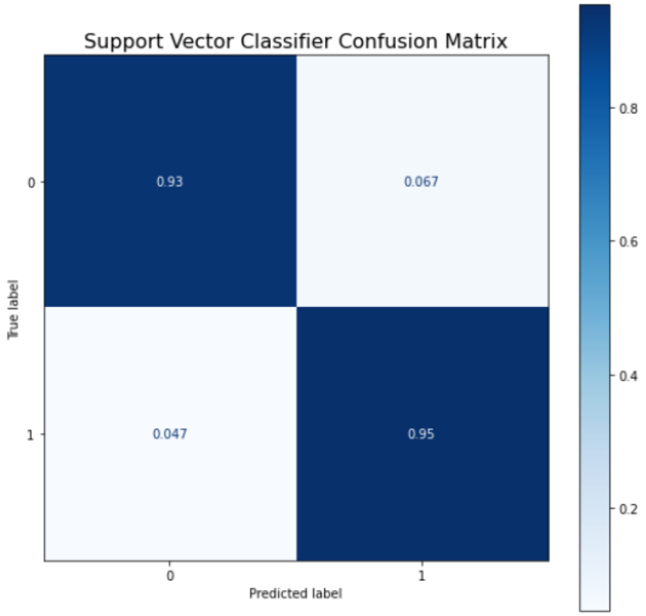
	Vectorizer	Classifier	Train Accuracy Score	Test Accuracy Score	ROC-AUC
0	TfidfVectorizer()	MultinomialNB()	1.0	0.958106	0.959476
1	TfidfVectorizer()	SVC(random_state=42)	1.0	0.943534	0.942701

Best Models

Multinomial NB vs SVM

CONFUSION MATRIX



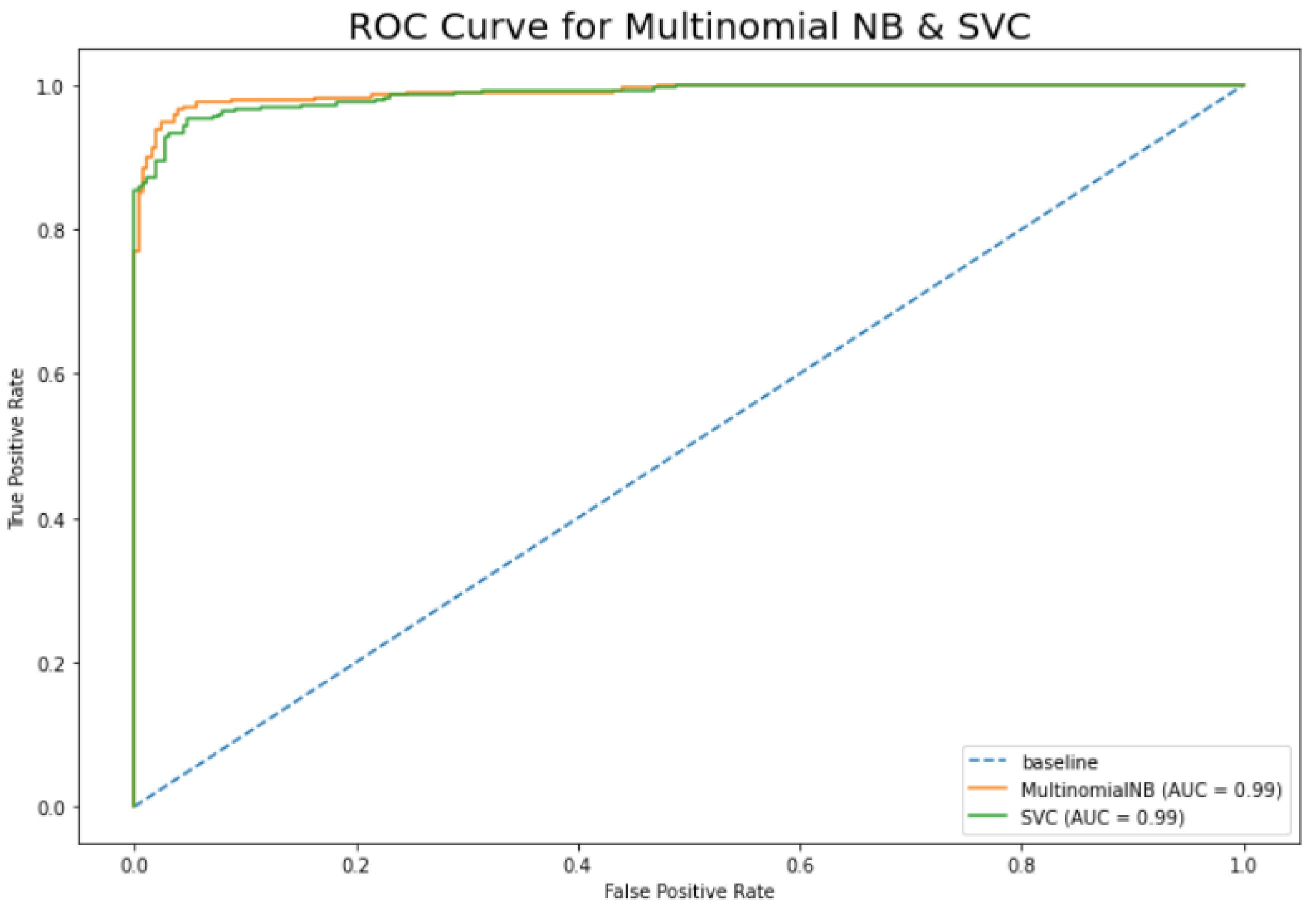


MNB better at correctly predicting posts from r/AskScience

SVM slightly better at correctly predicting posts from r/AskSocialScience

Multinomial NB vs SVM

ROC CURVE



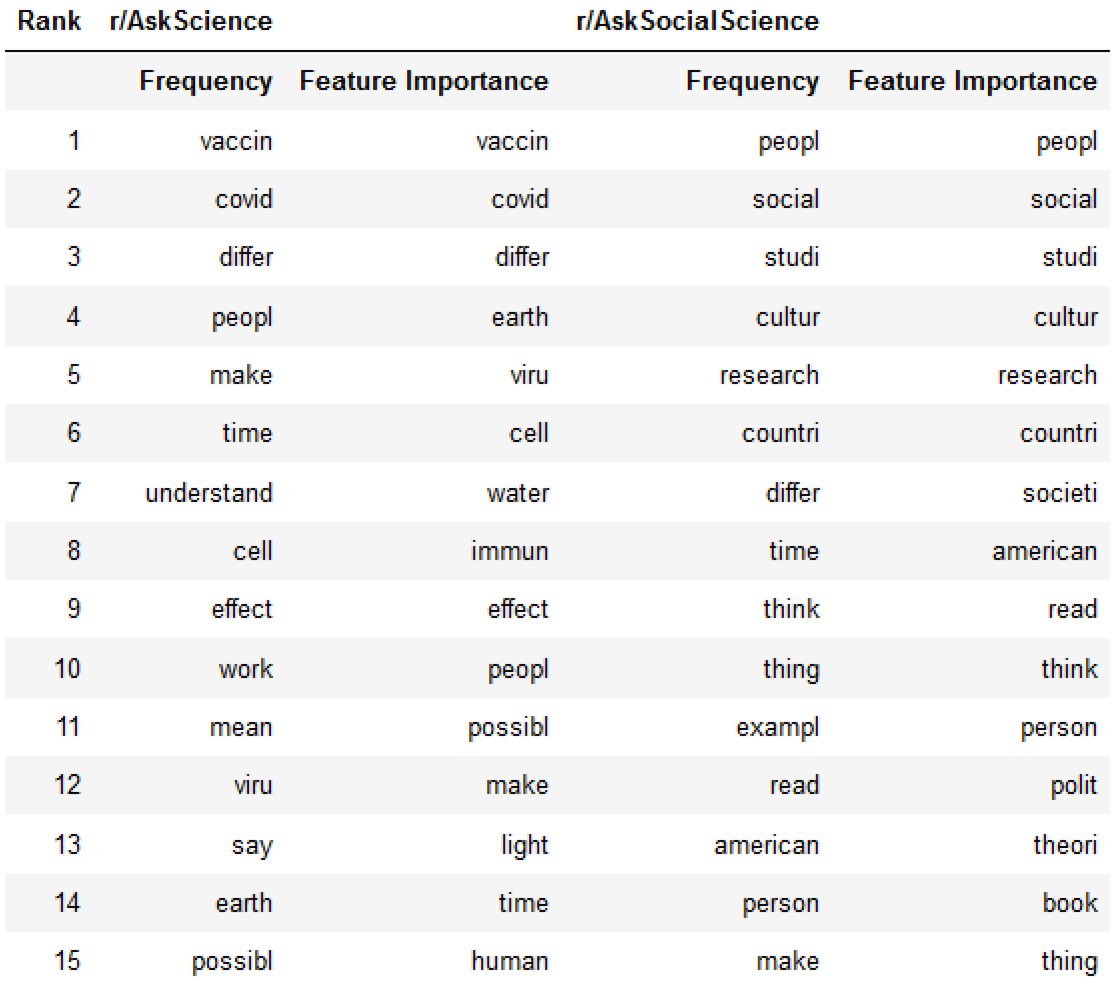
Multinomial NB performs slightly better

Multinomial NB vs SVM

INTERPRETABILITY

Multinomial NB

Able to extract feature importance



Rank	r/AskScience		r/AskSocialScience	
	Frequency	Feature Importance	Frequency	Feature Importance
1	vaccin	vaccin	peopl	peopl
2	covid	covid	social	social
3	differ	differ	studi	studi
4	peopl	earth	cultur	cultur
5	make	viru	research	research
6	time	cell	countri	countri
7	understand	water	differ	societi
8	cell	immun	time	american
9	effect	effect	think	read
10	work	peopl	thing	think
11	mean	possibl	exampl	person
12	viru	make	read	polit
13	say	light	american	theori
14	earth	time	person	book
15	possibl	human	make	thing

Multinomial NB vs SVM

INTERPRETABILITY

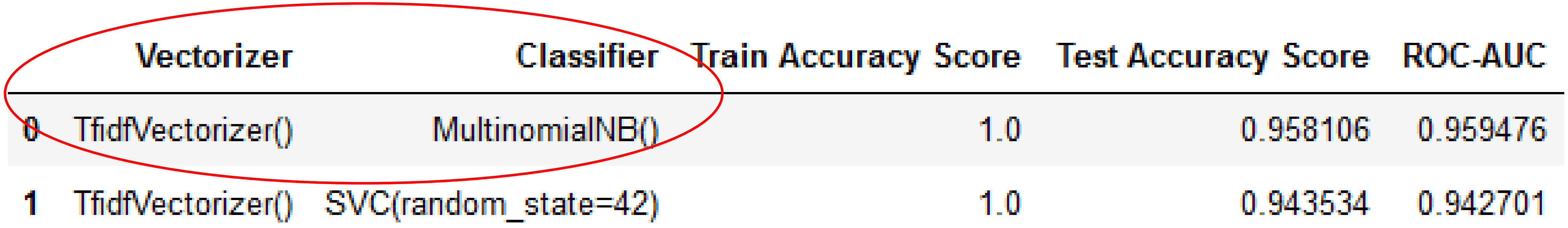
SVM

‘Black Box’ Model

Difficult to interpret coefficients

Linear SVC

- Coefficients represent vector coordinates
  - Orthogonal to hyperplane which separates classes
- Take dot product with new observation point
  - If positive, classify as positive class
  - Vice versa
- Importance of feature can be estimated
  - Absolute size of coefficient relative to others



	Vectorizer	Classifier	Train Accuracy Score	Test Accuracy Score	ROC-AUC
0	TfidfVectorizer()	MultinomialNB()	1.0	0.958106	0.959476
1	TfidfVectorizer()	SVC(random_state=42)	1.0	0.943534	0.942701

Best Model

Model Limitations

Misclassified Posts

- ‘Close calls’ – probabilities close to 0.5
  - Why is the Cardia (oesophagus-stomach opening) named so?
  - I'm curious about the linguistics (?) and the reasoning behind whoever named that region, considering that everything heart-related is "cardiac," but just recently I learned that anything related to the Cardia is also "cardiac". They both seem to be from the Greek word "kardia" (heart) according to Merriam Webster, so I'm curious if something got lost in translation or if the scientist naming that region just decided to be funny.
- Model unable to interpret semantics of post
- Too many important features for both subreddits

Model Limitations

Misclassified Posts

- Completely wrong
  - Is climate change boosting development of mountainous regions and therefor of more mountainous countries?
  - Looking at constant decrease of snow got me wondering: Is proportion of tourists in areas where there's substantial amount of snow cover (mountainous and northern regions) increasing as a result? I for one would've loved to see Paris during winter but seeing that there's no snow made snowy areas more lucrative for me as I could visit Paris at any time of year and the experience wouldn't be very different. How do you think will that develop mountainous rural areas and countries which are covered with those areas to substantial amount, like Switzerland or Austria?
- Model unable to interpret semantics of post
- Question on social impact (development)
- Premise of question on climate change

Improvements

More data!!!

- Reddit API limitations
- Limited amount & timespan of data
- Alternative APIs (Pushshift API)

Incorporate semantic concepts into model

- Sentiment analysis
- Relationship Extraction

Conclusions & Recommendations

Effective classification model using NLP techniques

- Tfidf Vectorizer + Multinomial Bayes Classifier
- High predictive performance + useful for inference

Useful for subreddit moderators

- Inferences can increase understanding of underlying characteristics of community
  - Shape moderation policies & influence direction of subreddit
- Can help solidify identity of subreddit
  - Discover themes & important topics
  - Boost engagement with community