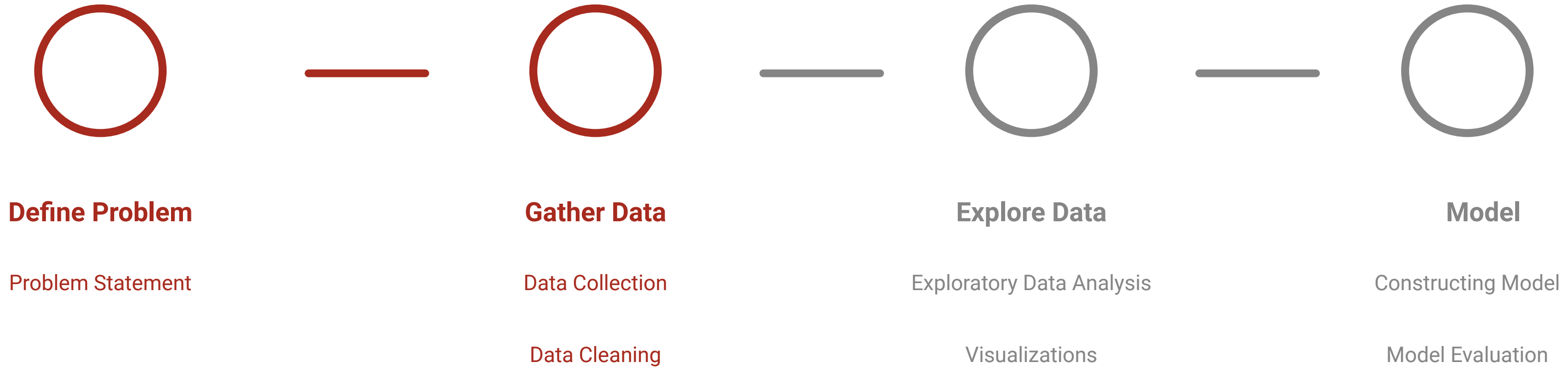
Subreddit Classification

Shandeep
DSI14

Table of Contents



Define Problem

Problem Statement

Gather Data

Data Collection

Data Cleaning

Explore Data

Exploratory Data Analysis

Visualizations

Model

Constructing Model

Model Evaluation

Define Problem

community cases
jumpstart account
year old
short term
outside dorms
code contribute
circuit breaker period
mental health
don think
press conference
enforcement officer
financial literacy
covid 19
don know
fixed deposit
thanks reply
mentally ill
circuit breaker
savings account
abdul halim
coronavirus outbreak
source code contribute
social distancing
breaker period
wisdom teeth
source code
ont router
dont know
10 000
distancing enforcement
sorry loss
safe distancing
place residence
11 look
advertisement advertisement
nuclear power
income tax
started new job
ve got
premier league
don want
year olds
ve heard
going work
long term
support staff
10 years
common sense
worst case
fossil fuels
does mean
furlough scheme
better job
secondary school
boris johnson
good job
self employed
don like
don think
triple lock
universal credit
think people
year old
covid 19
socially distance
stay alert
think going
mr sunak
social distancing
public sector
spend money
herd immunity
isn going
case scenario
poor people
teaching assistants

Problem Statement:

The aim of this project was to analyse the language used in these subreddits and to examine whether or not they were unique to both Singapore and the United Kingdom

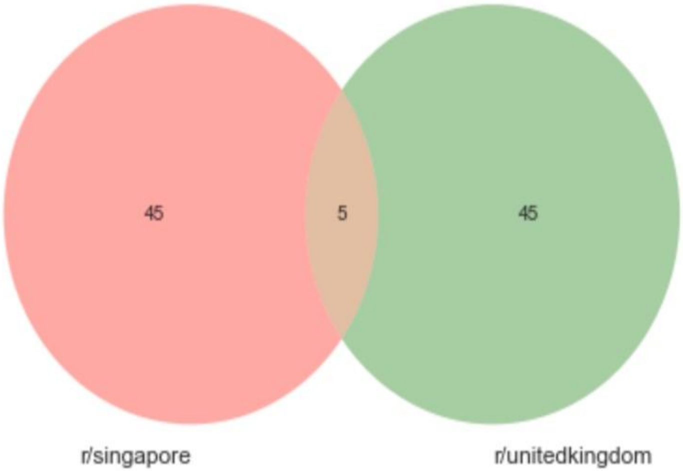
Gather Data

Data Collection: r/singapore & r/unitedkingdom

Number of data: 944 and 956 comments

X = comments, y = subreddit

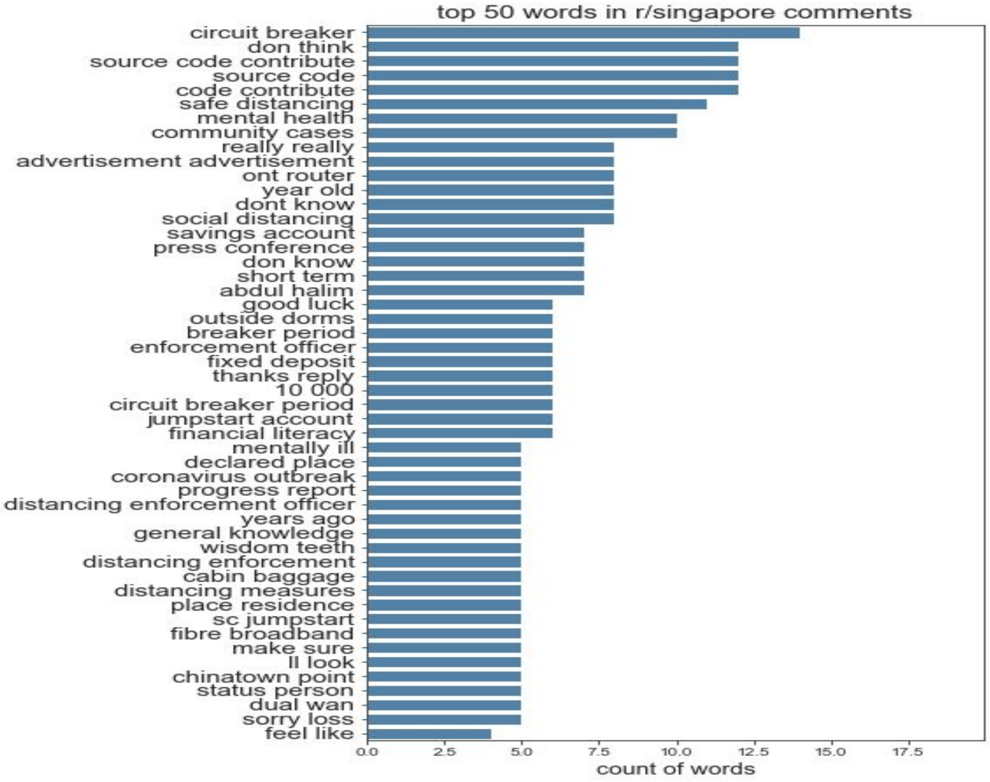
Explore Data

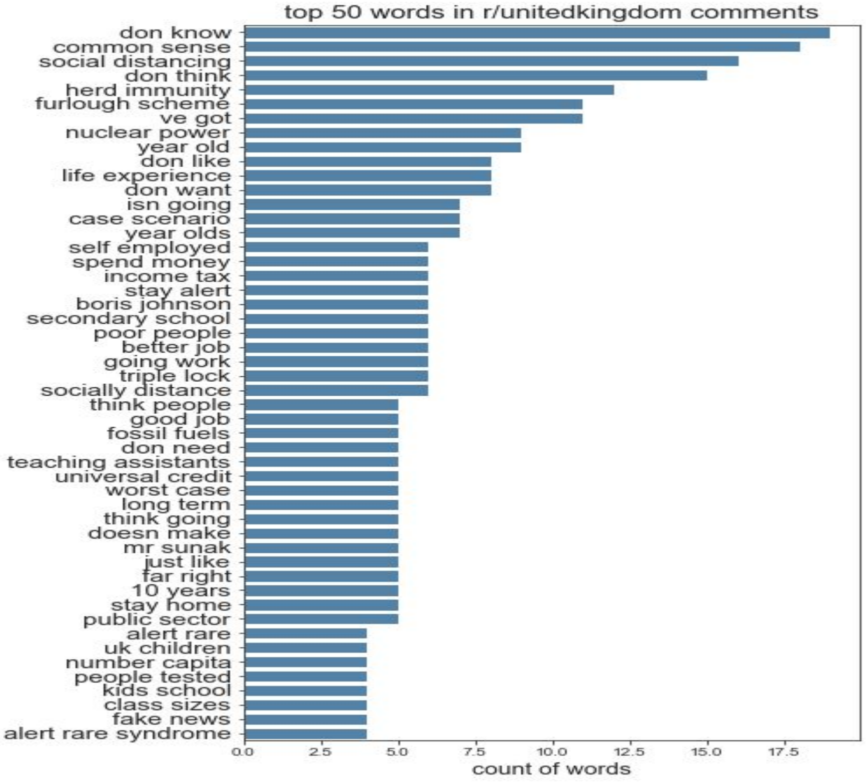


Exploratory Data Analysis:

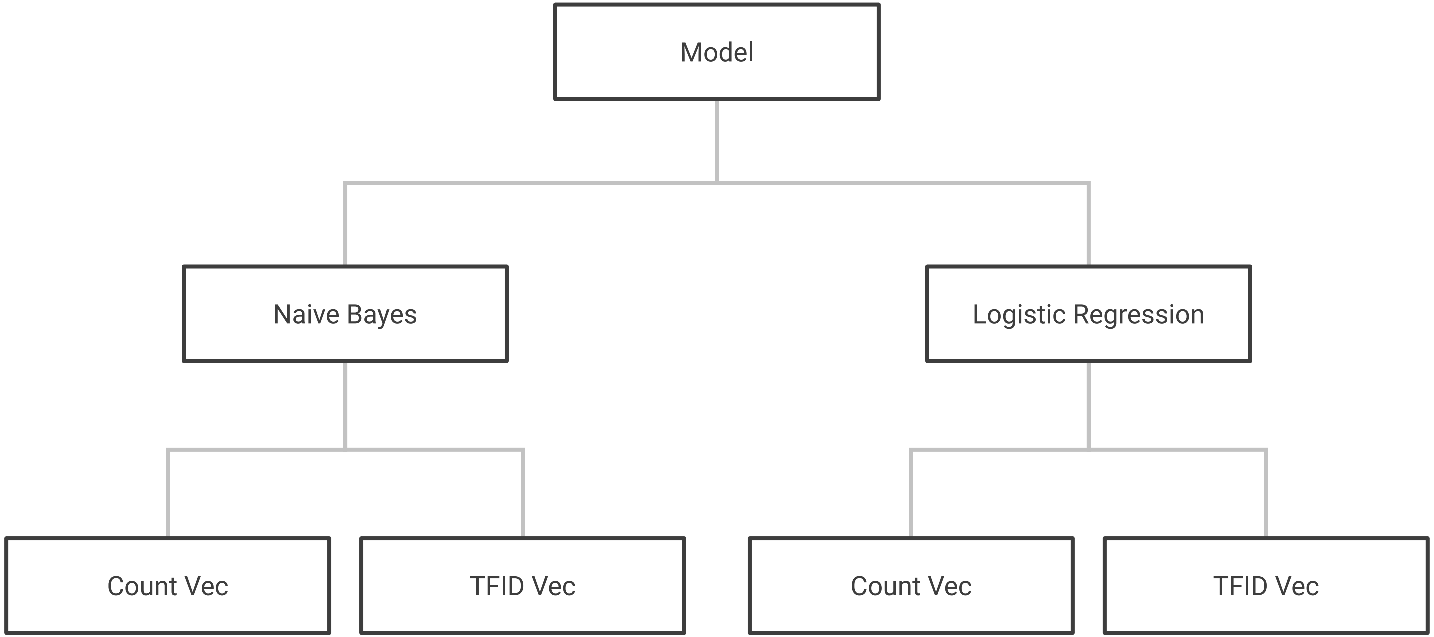
- “Covid 19” top word in /singapore, “Don know” top word in /unitedkingdom
- Covid related issues dominate both subreddits
- Only 5 out of top 50 phrases similar (2,3 ngram)
- **Average number of words in comments:** 39 for r/singapore and 45 for r/unitedkingdom

Explore Data





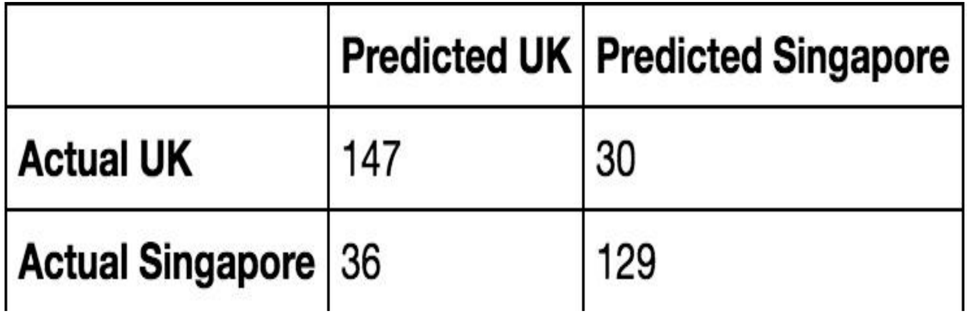
Model with Data



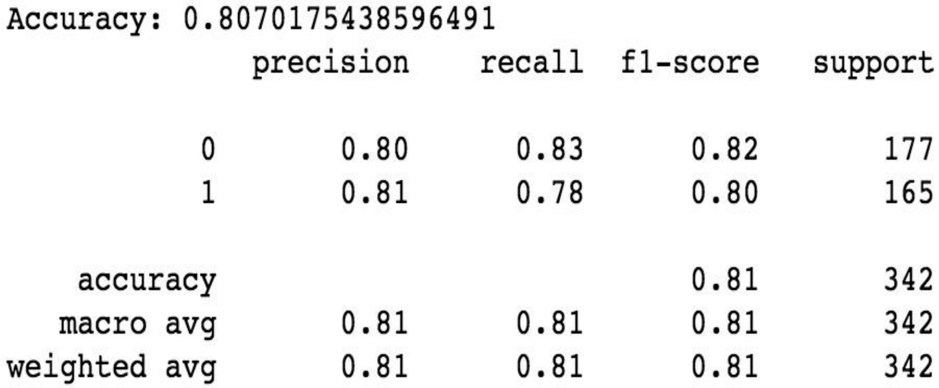
Modeling with Data

Model	Training Score	Test Score
CVEC + Naive Bayes	0.99	0.68
TFID + Naive Bayes	0.99	0.69
CVEC + Logistic Regression	0.99	0.76
TFID + Logistic Regression	1	0.80

Model Evaluation



	Predicted UK	Predicted Singapore
Actual UK	147	30
Actual Singapore	36	129



```
Accuracy: 0.8070175438596491
              precision    recall  f1-score   support

           0       0.80      0.83      0.82       177
           1       0.81      0.78      0.80       165

    accuracy                           0.81       342
   macro avg       0.81      0.81      0.81       342
weighted avg       0.81      0.81      0.81       342
```

Conclusion:

- Despite limitations, model has predictive value
- both subreddits have a lot of covid language even after stop words
- differences stem mainly from current affairs
- larger corpus which incorporates larger vocab could be useful