# Are you OK?

Project 3: NLP and web scraping

Updated: 20 Jan 2022

# Introduction & Problem Statement

## Increase in mental health issues

Studies conducted by Institute of Mental Health and University of Hong Kong revealed that 13% of 1000 surveyed local participants reported anxiety or depression

## Early detection

Early treatment is key for successful recovery e.g. medical treatment or social support.

Neglected cases could lead to suicidal tendency.

## NLP classifier for detection

Explore the use of NLP classifier as a tool to identify troubled individuals through their social media posting.

## Training data set for NLP classifier

No training data set readily available

Novel approach, scraping reddit pots

Subreddit chosen: r/MentalHealthUK and r/britishproblems

# Why these subreddits

## Relevancy

- r/MentalHealthUK provides a pool of relevant text that is related to poor mental well-being
- r/britishproblems provides a pool of text as a baseline with negative sentiment.

r/MentalHealthUK
" I don't feel human, I feel more like an alien in my day to to day cause of autism. and none of it seems to matter to anyone. just a wee vent hopefully someone will read this"

r/bristishproblems
"Seeing women consistently being overlooked or held back with their careers by employers for going on maternity leave"

# Data Cleaning

Remove duplicates

Remove URL links

Remove subreddit keyword from text

# EDA

## N-grams exploration

Explore unigram, bigram, and trigram of each subreddits

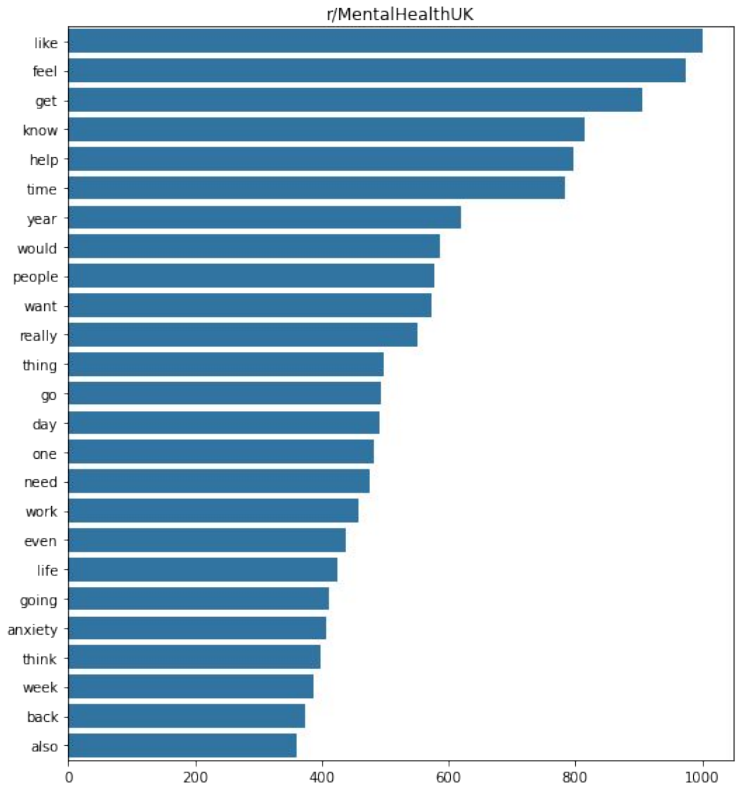
## Sentiment analysis

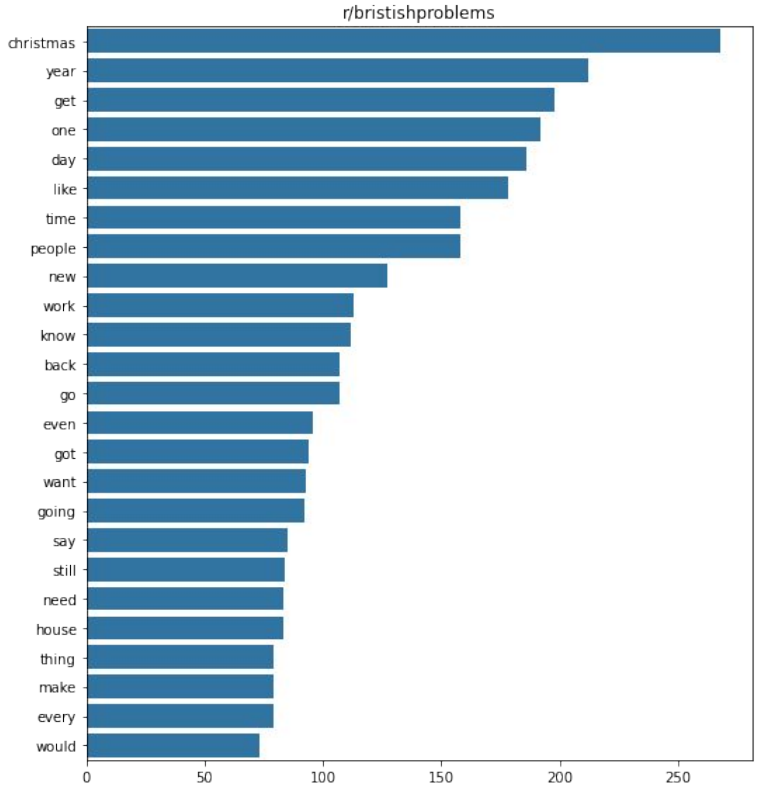
Investigate the average sentiment of each subreddit

## Post length

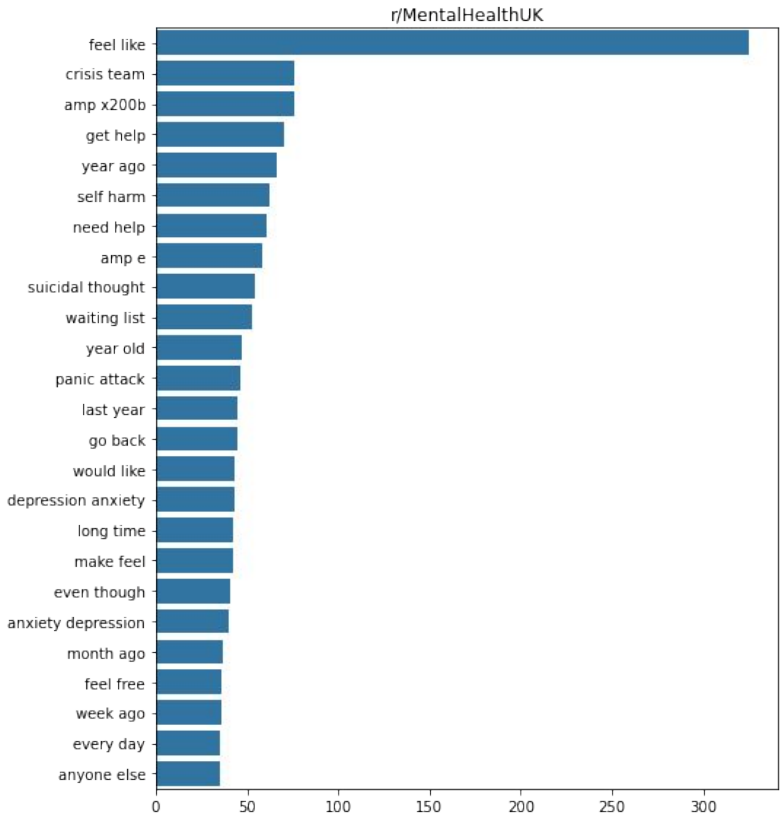
Investigate the statistic and distribution of each subreddit

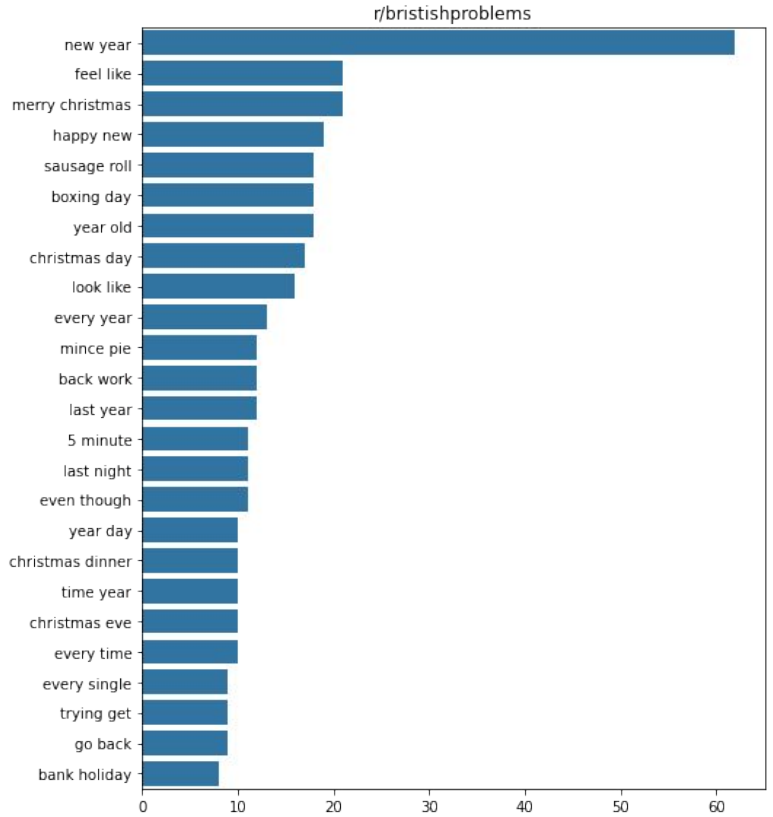
# Unigram


r/MentalHealthUK
like
feel
get
know
help
time
year
would
people
want
really
thing
go
day
one
need
work
even
life
going
anxiety
think
week
back
also
0
200
400
600
800
1000

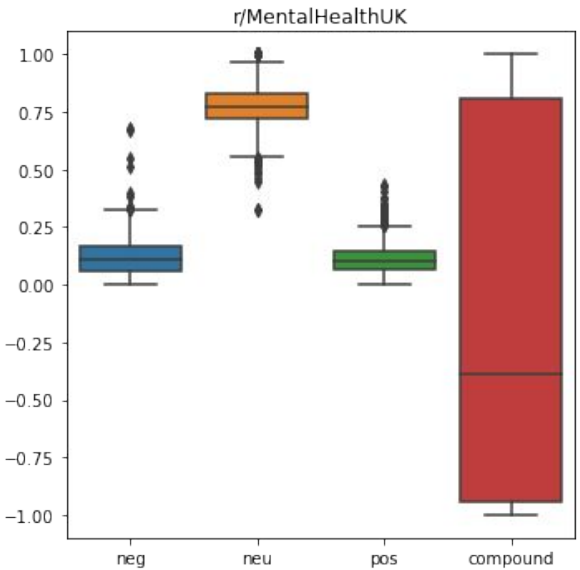

r/bristishproblems
christmas
year
get
one
day
like
time
people
new
work
know
back
go
even
got
want
going
say
still
need
house
thing
make
every
would
0
50
100
150
200
250

# Bigram

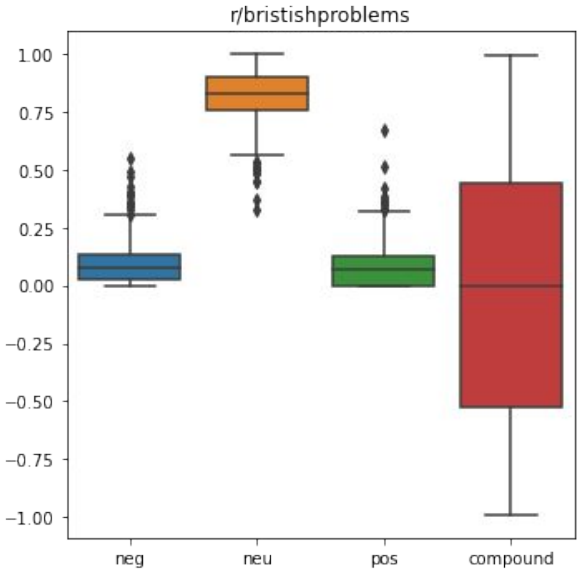

r/MentalHealthUK
feel like
crisis team
amp x200b
get help
year ago
self harm
need help
amp e
suicidal thought
waiting list
year old
panic attack
last year
go back
would like
depression anxiety
long time
make feel
even though
anxiety depression
month ago
feel free
week ago
every day
anyone else
0
50
100
150
200
250
300


r/bristishproblems
new year
feel like
merry christmas
happy new
sausage roll
boxing day
year old
christmas day
look like
every year
mince pie
back work
last year
5 minute
last night
even though
year day
christmas dinner
time year
christmas eve
every time
every single
trying get
go back
bank holiday
0
10
20
30
40
50
60

# Sentiment analysis




**median compound = - 0.4**

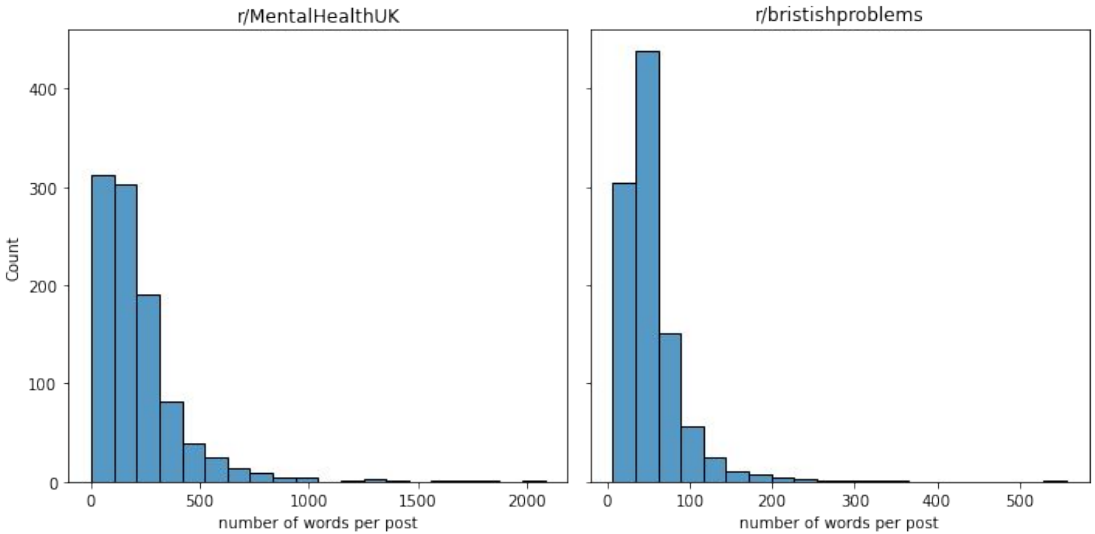



**median compound = 0**

## VADER sentiment scores

- The pos, neu, and neg scores are ratios for proportions of text that fall in each category.
- positive sentiment: compound score >= 0.05
- neutral sentiment: -0.05 < compound score < 0.05
- negative sentiment: compound score <= -0.05

# Post length




## # words per post

- Both subreddits have right skewed distribution
- 4 times more words per post in r/MentalHealthUK compared to r/britishproblems

# Modelling

## Bag of words method

Bag-of-words method treat text as a collection of individual words

Frequency of occurrence of each word is used as a predictor

Alternative, word sequence method. It place importance on word order

## Vectorization, lemmatization

String of text is tokenized into word before converted into sparse matrix

Change words into its root form
e.g playing -> play.

Reduce the words variation due to grammar rules

## Metric for evaluation

Envisage to identify as many potential individuals w/ mental health issues. As such, model will be built to **optimise the recall** performance.

In other words, the model is expected to predict lesser false negative but with a trade off of more of false positive.

## Classification algorithms

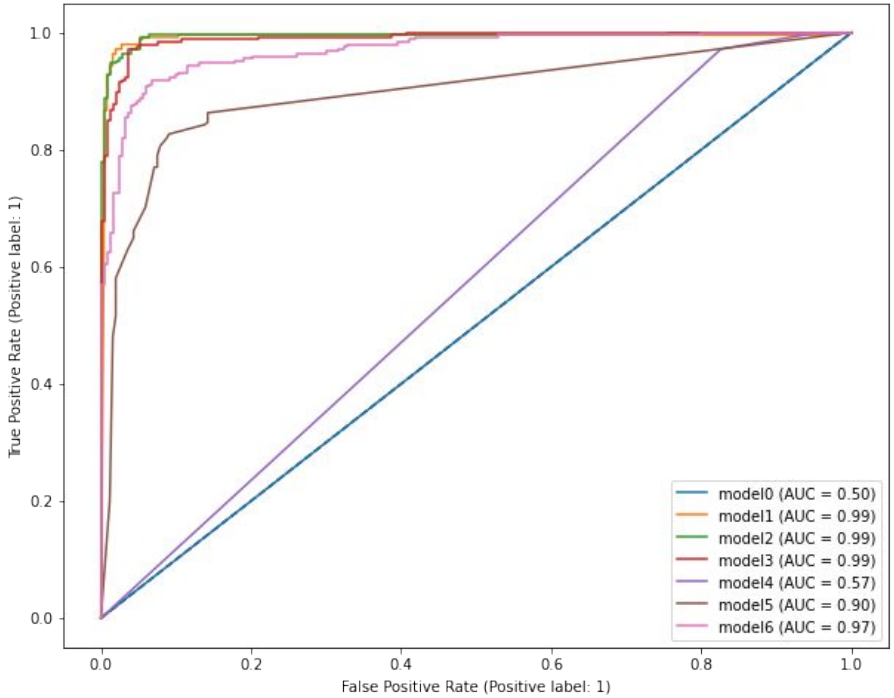
1. Dummy Classifier (baseline)
2. Multinominal Navie Bayes
3. Logistic Regression
4. K-nearest Neighbour
5. Decision Tree Classifier
6. Random Forest Classifier

hyperparameters optimized via grid search

# Results

| Models | Accuracy | Precision | Recall | Specificity | F1 | ROC AUC |
|---|---|---|---|---|---|---|
| Model0: Dummy classifier | 0.469 | 0.465 | 0.480 | 0.458 | 0.472 | 0.5 |
| Model1: count vectorizer, multinomial NB | 0.928 | 0.876 | 0.996 | 0.862 | 0.932 | 0.99 |
| Model2: TFIDF, multinomial NB | 0.906 | 0.843 | 0.996 | 0.818 | 0.913 | 0.99 |
| Model3: TFIDF, logistic regression | 0.946 | 0.966 | 0.923 | 0.968 | 0.944 | 0.99 |
| Model4: TFIDF, K-nearest neighbour | 0.517 | 0.506 | 1.000 | 0.0435 | 0.672 | 0.57 |
| Model5: TFIDF, decision tree | 0.864 | 0.909 | 0.806 | 0.921 | 0.855 | 0.90 |
| Model6: TFIDF, random forest | 0.918 | 0.956 | 0.875 | 0.921 | 0.914 | 0.97 |

# ROC Curve




| Models |
| --- |
| Model0: Dummy classifier |
| Model1: count vectorizer, multinomial NB |
| Model2: TFIDF, multinomial NB |
| Model3: TFIDF, logistic regression |
| Model4: TFIDF, K-nearest neighbour |
| Model5: TFIDF, decision tree |
| Model6: TFIDF, random forest |

# Conclusion

## Production model

- TFIDF, multinominal NB
- Recall : 0.996
- ROC AUC: 0.990
- Yes, NLP classifier is a viable early detection tool to assist healthcare institutes, government bodies

## words association

- important features (words) identified by model
- word cloud on next slide

## Improvement

- To improve model further, the training set should consider seasonality cycle. i.e., sampling from Jan - Dec could be a better approach.
- Word sequence method can be used instead of bag-of-word method
- Subreddits chosen based on UK community. As such there might be some linguistic differences


help
get
feel
year
know
like
time
people
would
anxiety
work
really
want
need
think
go
going
month
anyone
friend
told
day
life
therapy
support
nh
please
another
trying
issue
depression
suicidal
assessment
service
care
job
feeling
something
week
Said
psychiatrist
anything
find
question
treatment
therapist
quite
also
much
every
made
taking
talking
actually
worse
ask
life
experience
crisis
thought
could
one
self
referred
two
sure
without
started
different
getting
private
school
home
thanks
hard
sertraline
diagnosis
past
hope
live
cbt
hi
someone
got
last
say
called
adhd
currently
hospital
idea
sleep
advice
team
good
see
since
keep
parent
problem
went
better
make
still
go
time
people
christmas
back
year
car
every
new
kid
minute
time
fucking
house
get
work
home
box
people
tea
way
xmas
watching
week
come
hour
price
need
full
chocolate
last
thing
family
shop
asked
eve
3
getting
going
say
know
born
tomorrow
instead
street
bit
someone
even
got
outside
around
buying
night
think
give
make
well
dinner
want
parent
wife
merry
delivery
driver
online
buy
dog
would
look
thought
card
still
door
half
saying
go
shit
eat
coffee
bed
knowing
2022

# Thank you