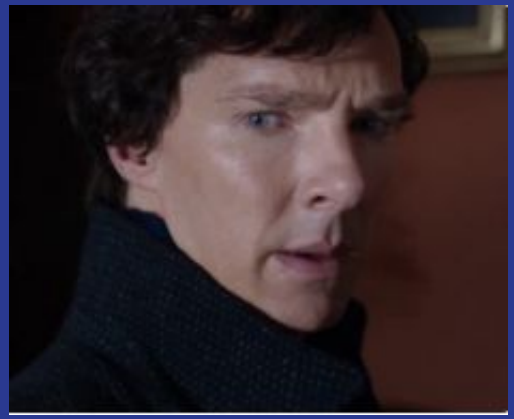
Elementary,

my dear Sherlock



Pictures from https://www.tvguide.com/

Problem Statement

The television shows 'Elementary' and 'Sherlock' are about the detective Sherlock Holmes from the studios CBS and BBC respectively. The CBS executives want to use a model to classify the subreddit threads r/elementary and r/sherlock to determine whether in viewers' minds there are differences between 'Elementary' and the other show.

This project will develop a model to distinguish between the 2 subreddits. The positive class will be the posts from the elementary subreddit and the negative class will be sherlock subreddit posts. Logisitic Regression and Naive Bayes models will be used to classify the 2 subreddits. The accuracy scores will be used to evaluate and find the best model for classification.

The project will also analyse which words were used by the model for successful classification to help CBS executives find out which terms/topics resonate in viewers' minds to ensure the popularity of the show.

Data acquisition

Data scrapped from:

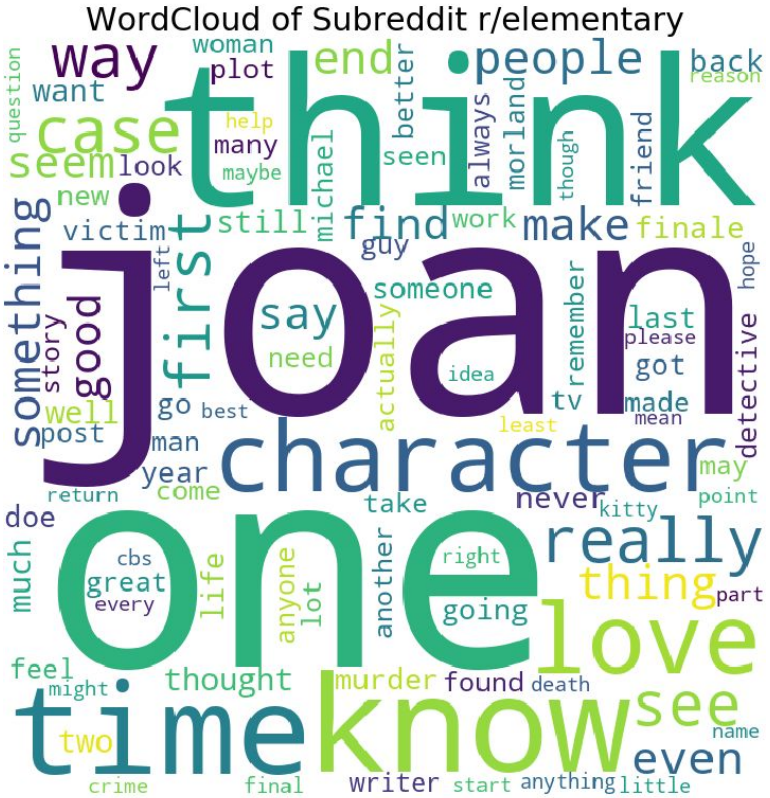
1) https://www.reddit.com/r/Sherlock/new

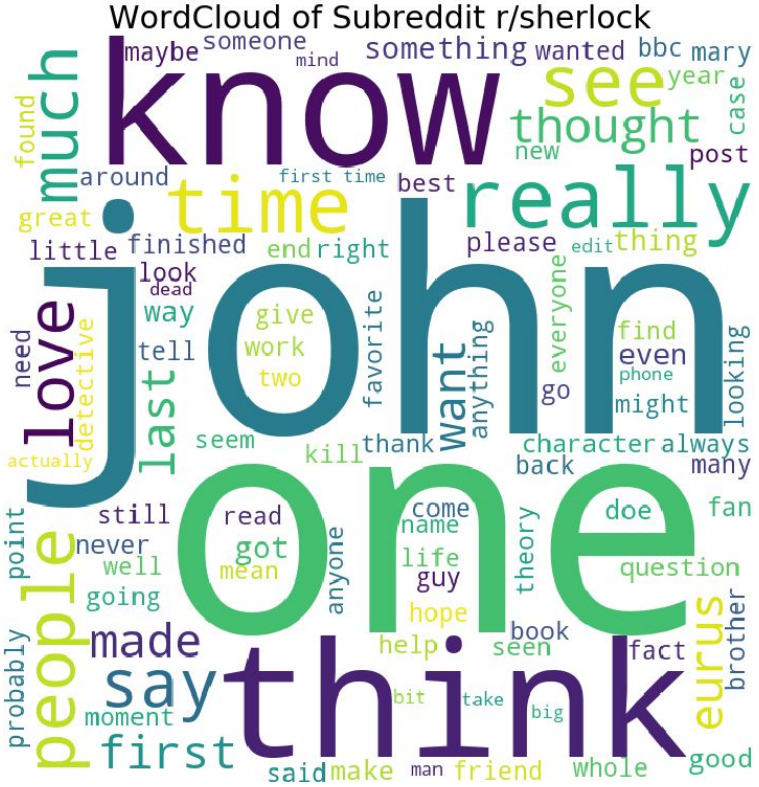
2) https://www.reddit.com/r/elementary/new

Data Cleaning

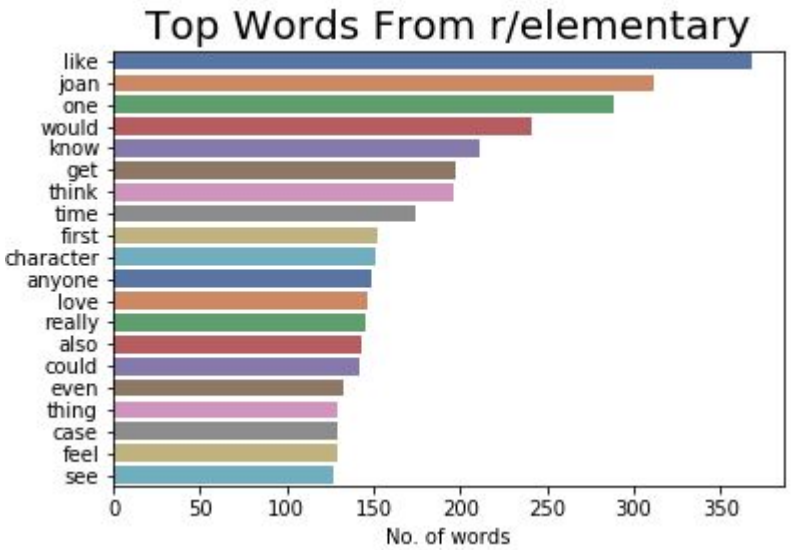
1. Null values to space
2. Duplicate rows dropped
3. Non -info columns dropped
4. Lemmatization
5. Remove urls
6. Remove common word
7. lower case

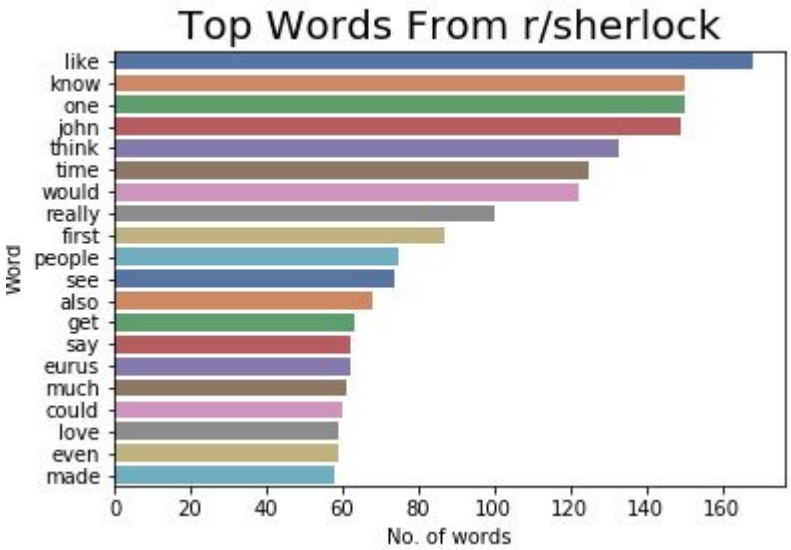
EDA


WordCloud of Subreddit r/elementary
joan
think
one
character
know
time
love
really
first
people
way
case
seem
something
good
find
make
end
back
thing
see
even
feel
thought
murder
found
writer
start
anything
little
final
crime
two
life
great
another
going
right
take
never
year
come
return
doe
much
cbs
every
anyone
lot
part
point
kitty
detective
hope
remember
tv
made
mean
actually
idea
least
post
man
go
best
well
story
left
question
woman
plot
want
help
many
maybe
look
new
victim
still
michael
seen
always
morland
though
friend
better
reason
finale
guy
someone
say
need
last
please
got

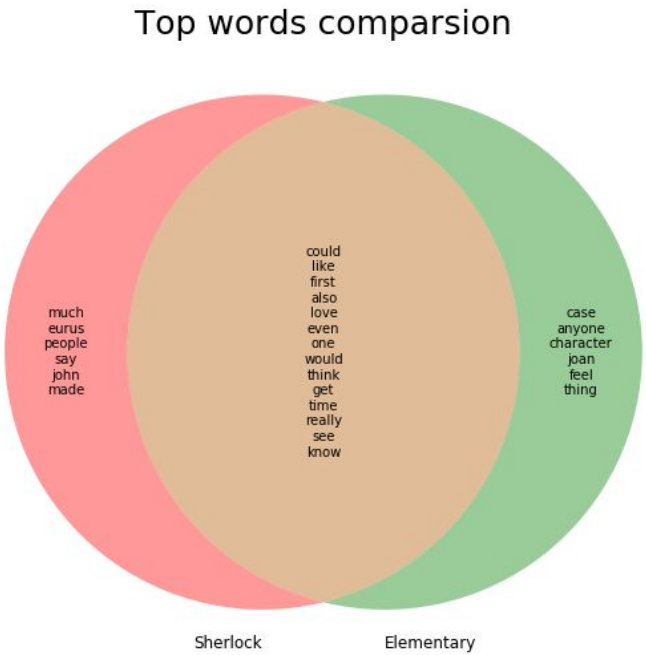

WordCloud of Subreddit r/sherlock
john
know
one
think
time
really
love
people
much
see
thought
last
want
made
say
first
eurus
maybe
someone
something
wanted
bbc
mary
year
case
post
new
around
first time
best
great
little
finished
end
right
please
edit
thing
look
dead
way
give
work
two
tell
favorite
everyone
find
even
phone
might
looking
anything
go
seem
thank
kill
character
always
back
many
actually
need
detective
still
read
come
name
never
got
mean
life
guy
theory
question
well
going
hope
book
help
seen
fact
brother
probably
moment
bit
take
big
said
make
man
friend
whole
good
point
anyone
doe
fan

EDA


Top Words From r/elementary
like
joan
one
would
know
get
think
time
first
character
anyone
love
really
also
could
even
thing
case
feel
see
0
50
100
150
200
250
300
350
No. of words


Top Words From r/sherlock
Word
like
know
one
john
think
time
would
really
first
people
see
also
get
say
eurus
much
could
love
even
made
0
20
40
60
80
100
120
140
160
No. of words

EDA


Top words comparsion
much
eurus
people
say
john
made
could
like
first
also
love
even
one
would
think
get
time
really
see
know
case
anyone
character
joan
feel
thing
Sherlock
Elementary

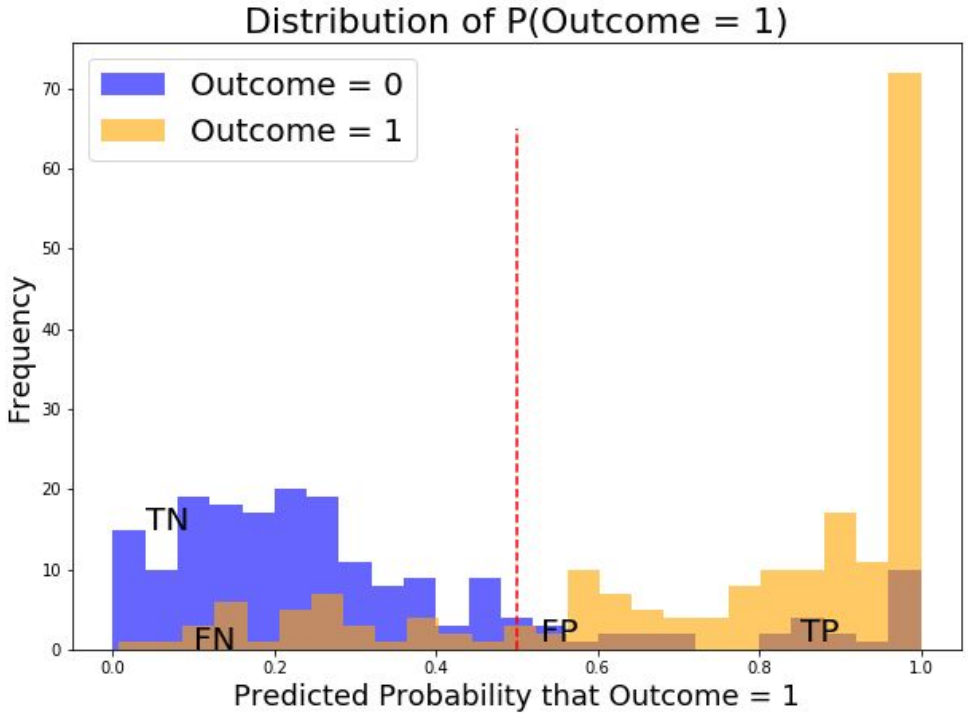
Model and Evaluation

	Logistic Regression	Multinomial Naive Bayes
Training Accuracy Score	0.9684	0.8661
Test Accuracy Score	0.8303	0.7738

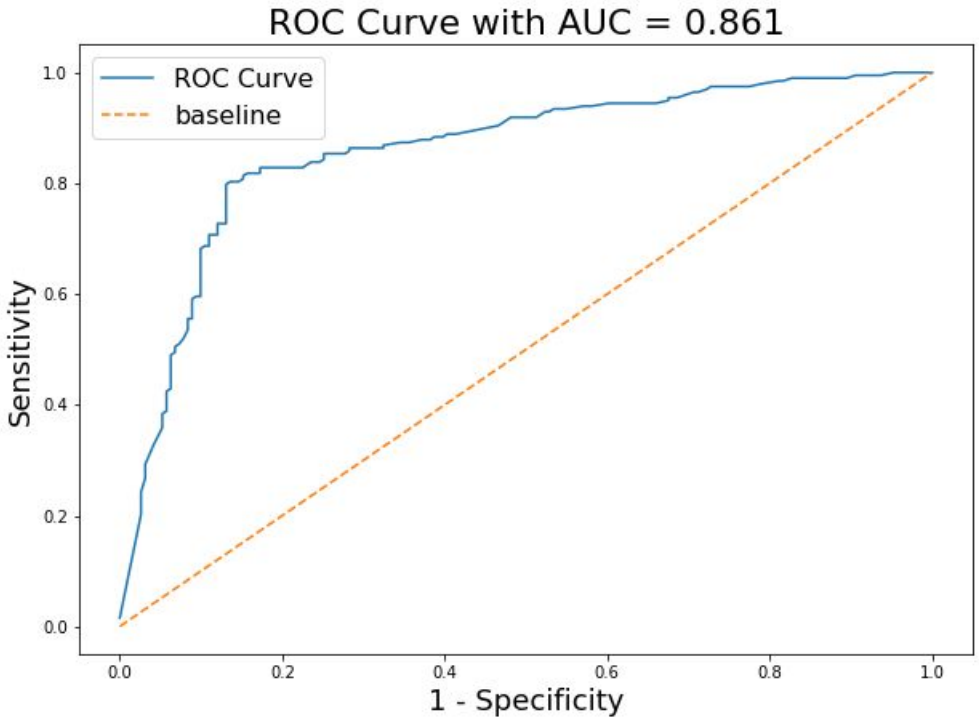
Model and Evaluation

Logistic Regression	
True Positive	162
True Negative	161
False Positive	30
False Negative	36
ROC AUC Score	0.8606

Model and Evaluation



Model and Evaluation



Model and Evaluation

	lemmatized	pred
546	factfilecan anyone find factfile need write full character report clue tip helpful thanks	1
1071	community activityso mod anything want gauge interest sub rewatch absolutely love played huge part life pretty inactive sub okay enough interest would like group sunday night site like rabbit disc...	0
128	anyone know discombobulated film used bbc sure cant find soundtrack	1
726	little sneak peek big meta writing put inside another huge meta s4basically analysis s4 stumble dialogue think enough get post posting mainly really proud bit really like whole concept got exited ...	1
1349	chicken	0

joan	2.029170
promo	1.842511
finale	1.489501
cbs	1.456896
renewed	1.433557
anyone	1.366018
return	1.249385
lucy	1.176168
hulu	1.148565
kitty	1.148551
miss	1.121215
sneak	1.097638
clyde	1.086568
discussion	1.058996
wear	1.052680
10	1.017161
actor	1.006725
bad	0.985270
gregson	0.974341
started	0.944931

Conclusion

1) Logistic Regression model selected
2) Model has a slight bias towards the negative class i.e r/sherlock posts.
3) Misclassified posts have non-relevant words
4) Both shows are distinct

Recommendation

1) Develop plots based on the the topics/themes found top 25 impactful words.

2) The model can be improved by:
   a) getting more data
   b) changing the hyperparameters
   c) reducing the no.of features
   d) using different classifiers and transformers.

Consideration

1) Data from most recent subreddit posts

2) The words used in both subreddits may be culturally different