

## Project.3

Unveiling The Recipe For  
Comedy Success:

Analyzing Viewer Preferences  
And Sentiment Towards Popular  
Sitcoms





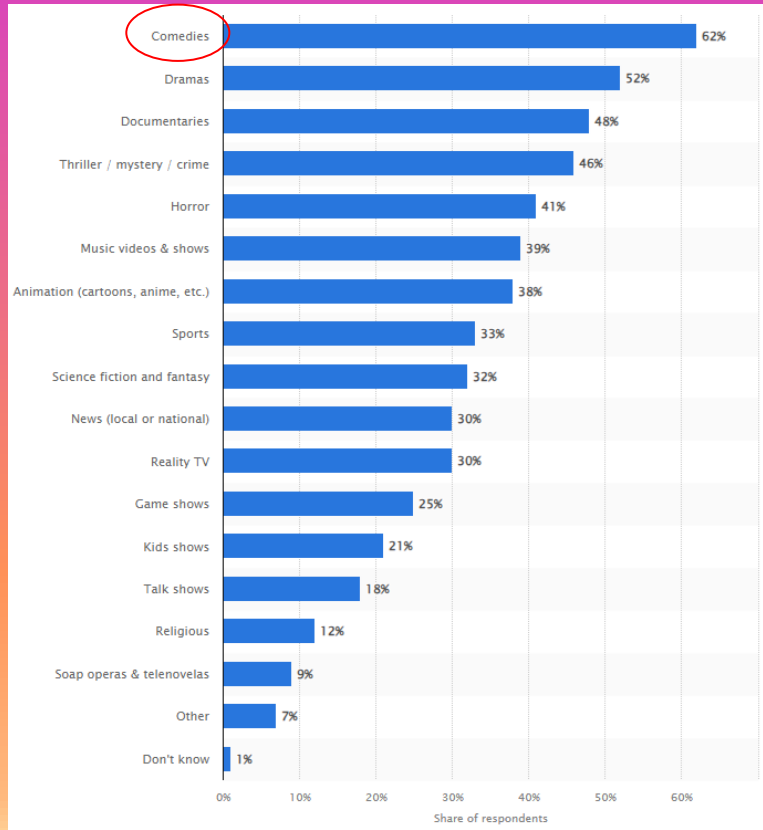
01

## Background Of The Problem



# 18 million

Average number of viewers  
tuning in to the last season  
of The Big Bang Theory



Preferred digital video  
content by genre in  
the U.S. as of March  
2023



# Streaming Services Market

01

## Netflix

223.09 million  
subscribers

02

## Prime Video

200+ million  
subscribers

03

## Disney+

164.2 million  
subscribers

04

## HBO Max

94.9 million  
subscribers



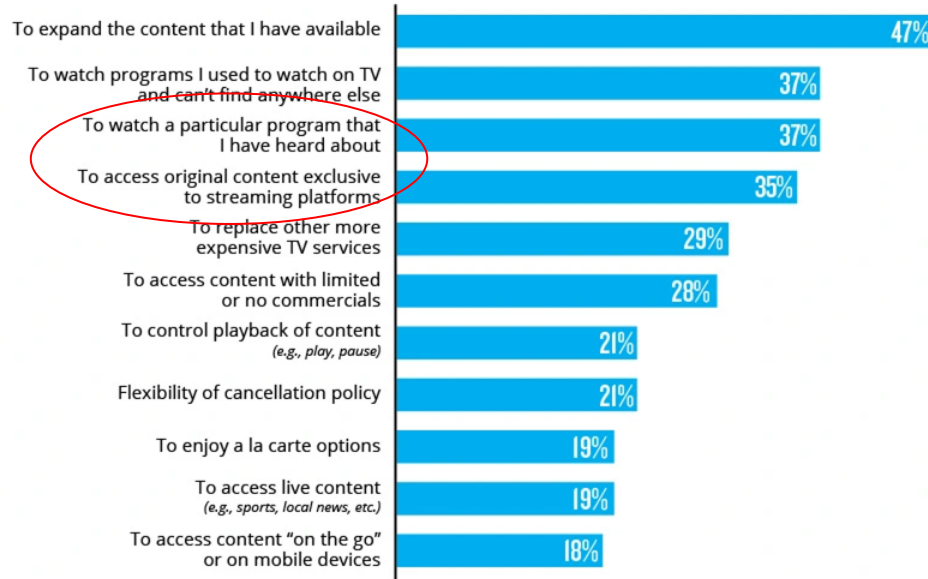
## IMPORTANCE OF VIDEO STREAMING ATTRIBUTES

Top 2 Box (out of 5 - Extremely/Very Important)



Availability of content is  
top 3 in importance of  
video streaming  
attributes

## REASONS FOR SUBSCRIBING TO ADDITIONAL PAID VIDEO STREAMING SERVICES



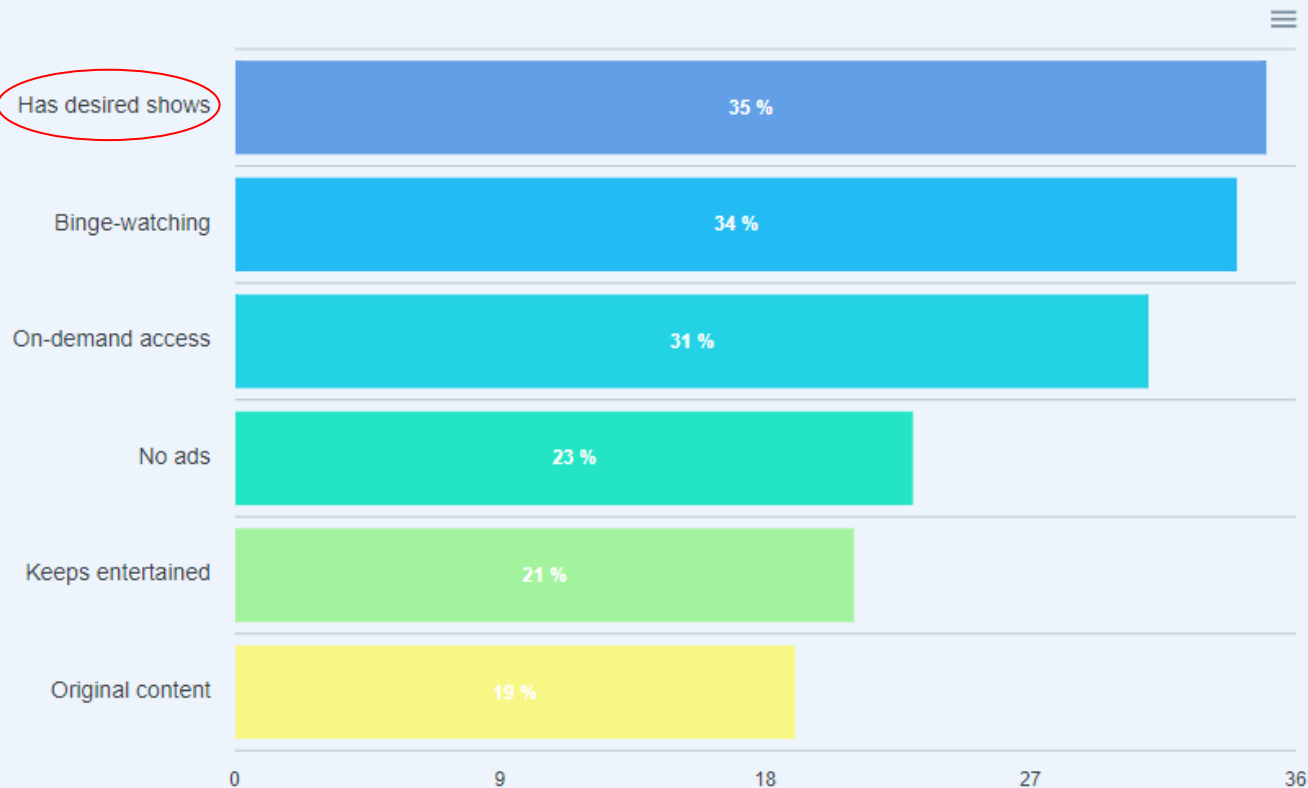
“Wanting to watch programs they’ve heard about”

&

“To access original content exclusive to streaming platforms”

Are cited as top 3 & 4 reasons for subscribing to additional paid video streaming services

## Most Attractive Features of Video Streaming Service

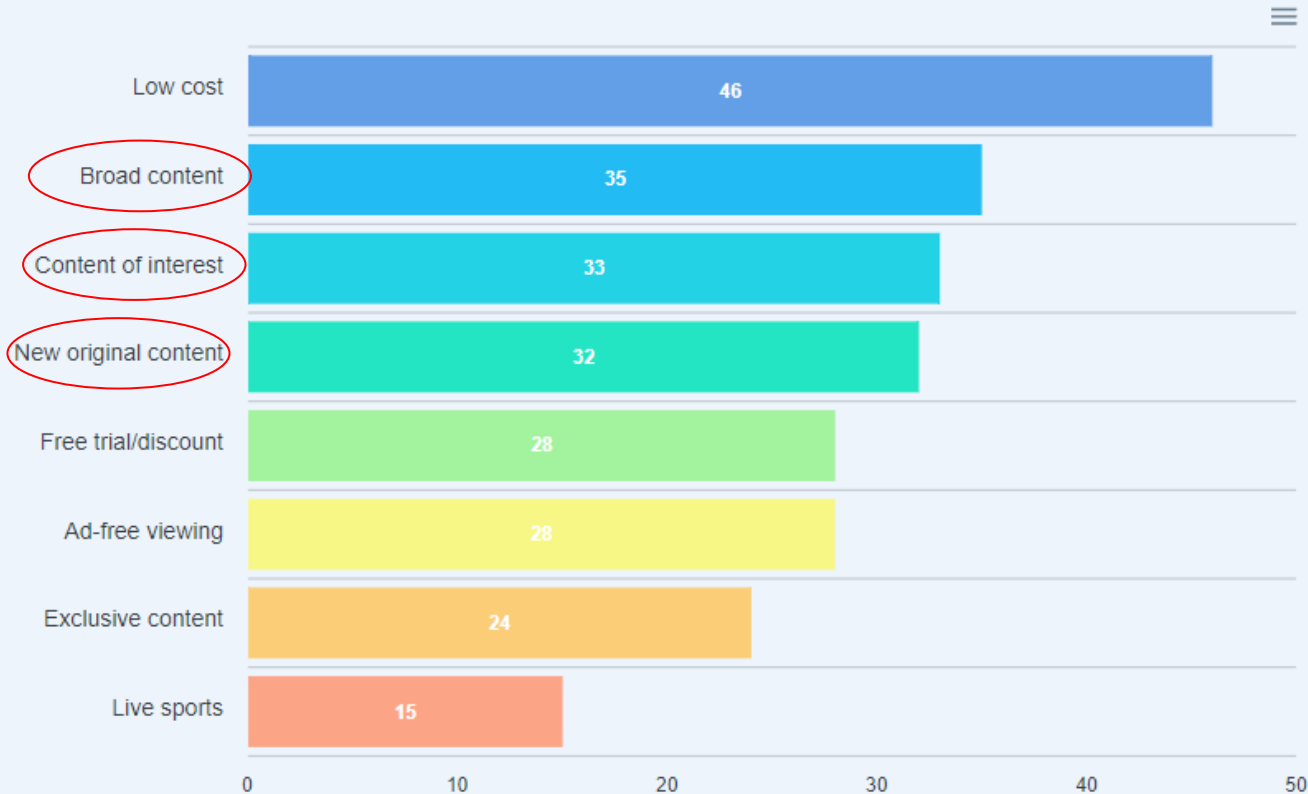


### Desired shows

Is cited as the top reason that consumers found to be the most attractive feature of video streaming services



## Top Reasons to Subscribe to a New Streaming Service

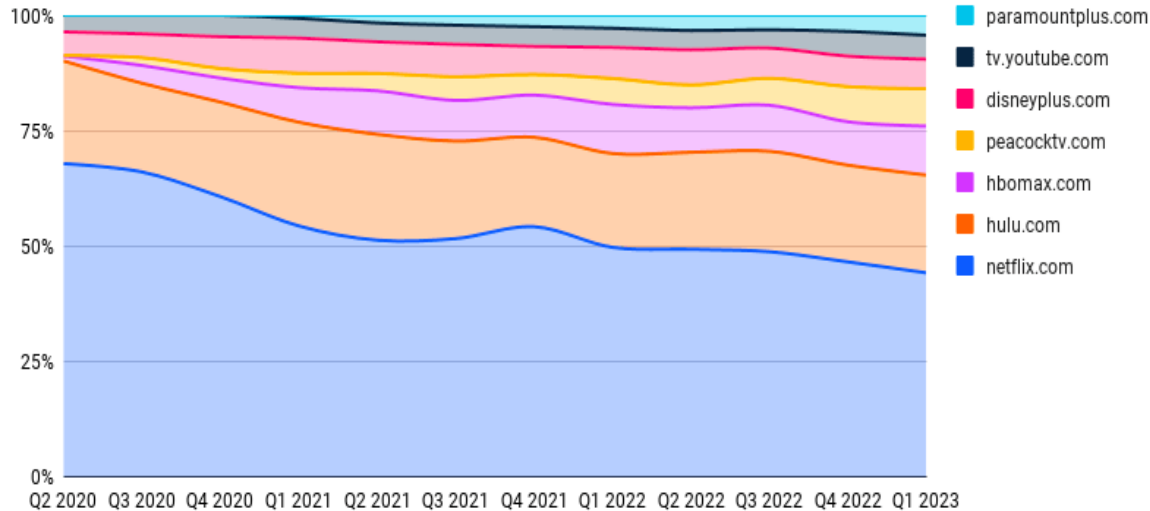


### Content related reasons

Are also cited as the top reason that incites consumers to subscribe to new streaming services

## Streaming Industry Digital Market Share

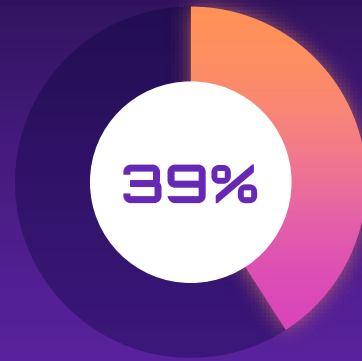
Traffic Share | Desktop & Mobile Web | March 2020 - March 2023



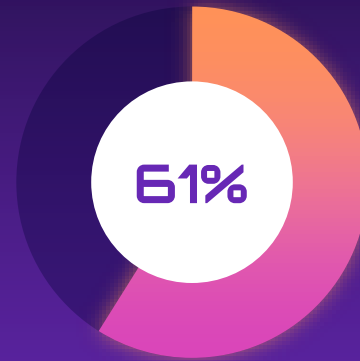
## SVoD industry competitiveness

Netflix has been losing market share every year from 2020 - 2023, reason being that there are more competition in the market.

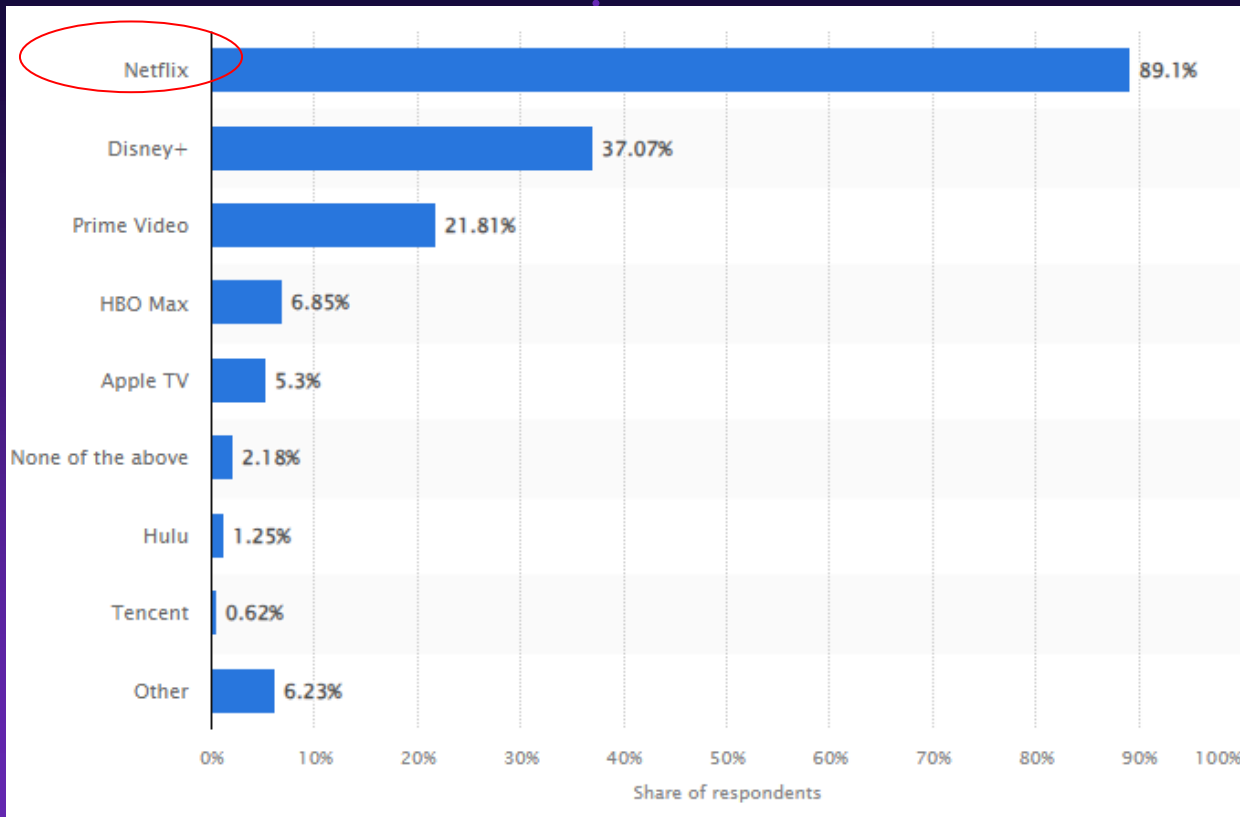
# When Singaporeans are asked if they would discontinue one of their streaming services in the next six months



Unlikely



Likely



**Most used streaming  
platforms among  
Singapore consumers in  
2022**

## Problem Statement:

### Short term problem

- a. Identify what show elements in the sitcom are popular among the viewers

01



### Brooklyn Nine-Nine

The team's favorite sitcom  
that inspired the team's name

02



### The Big Bang Theory

One of the all time best sitcom  
and also one of the team's  
favorites

## Problem Statement:

### Long term

- a. To build an infrastructure that can help to classify and analyse user's comments about the show from various platforms.

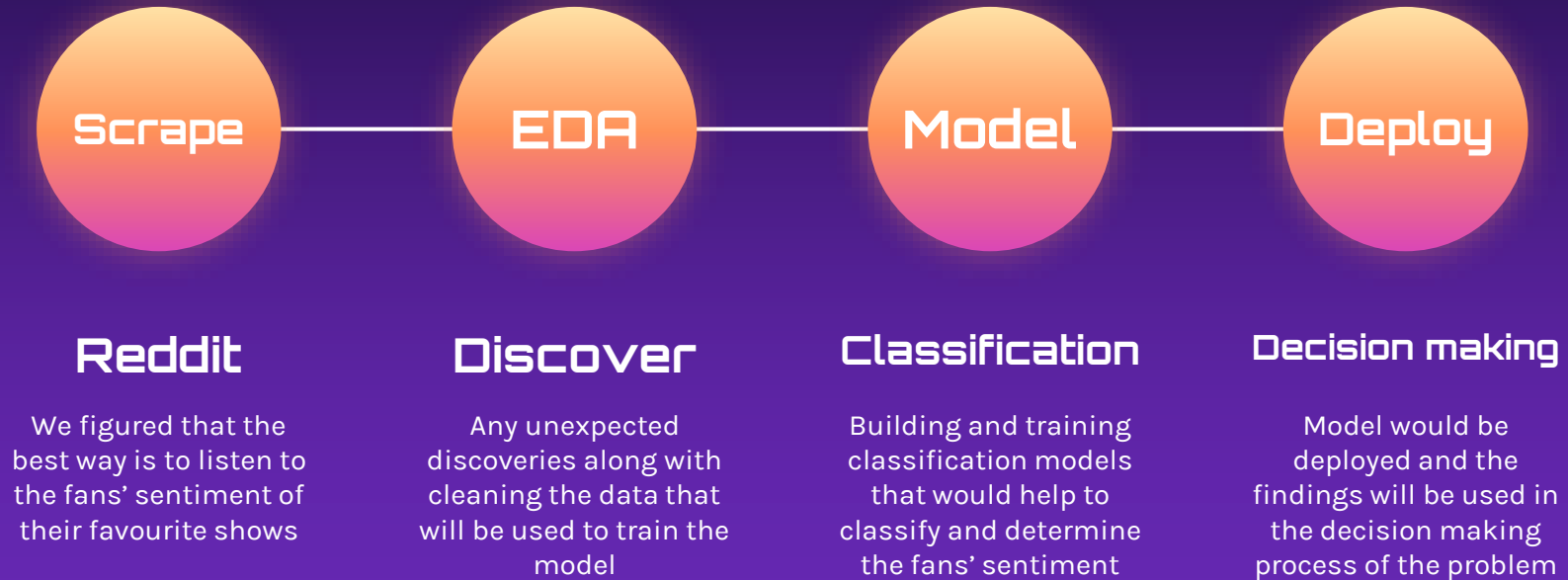
01

Classify comments by shows

02

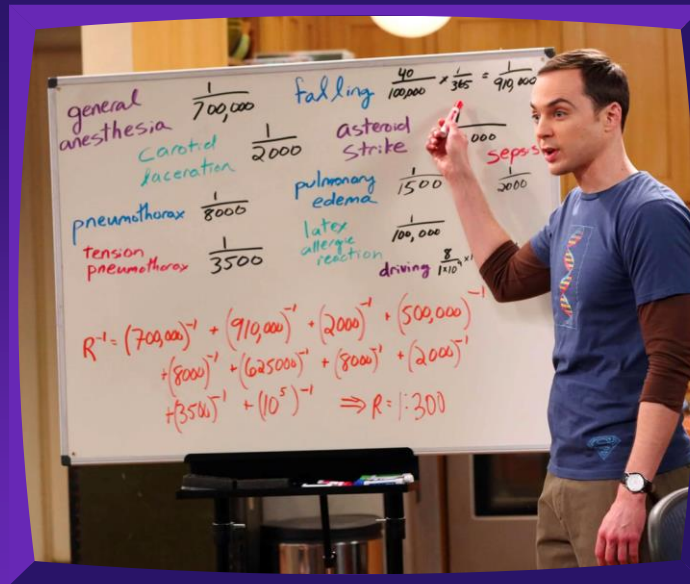
Analyse sentiments from the comments

# How we went about this



# 02

## Exploratory Data Analysis





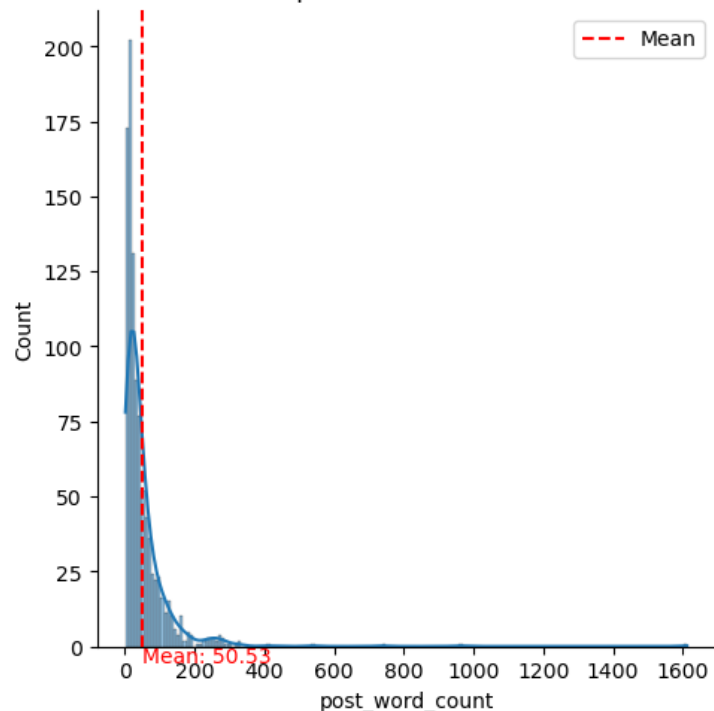
# Why Reddit?

- Free to use
- Freedom to express (unbiased input)
- Simple Accessibility of Data
- Large number of datasets (reduced biasness towards)
  - A genuine user reviews

# Word Counts per Post

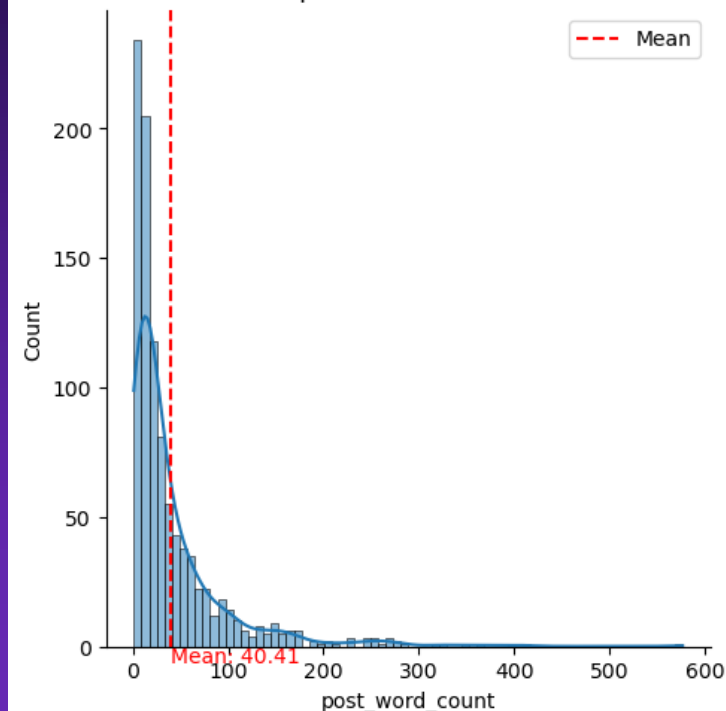
## r/bigbangtheory 50 words

Distribution of r/bigbangtheory  
posts word counts



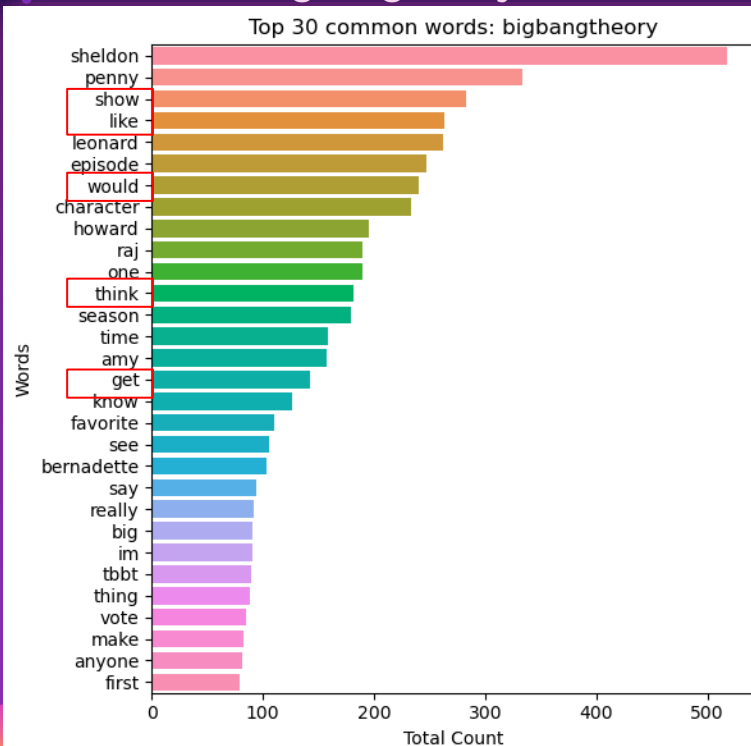
## r/brooklyninenine 40 words

Distribution of r/brooklyninenine  
posts word counts

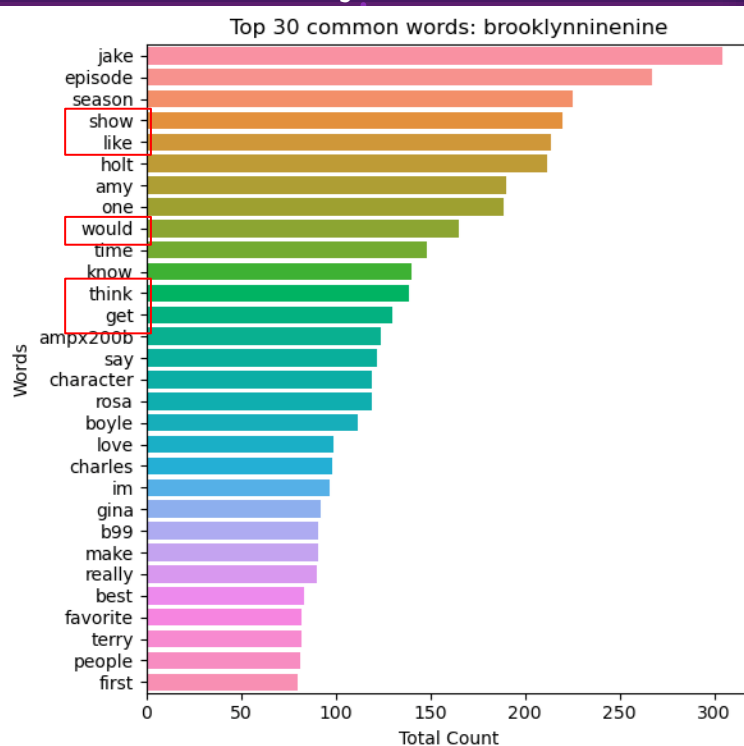


# Top 30 Common Words

r/bigbangtheory



r/brooklyninenine

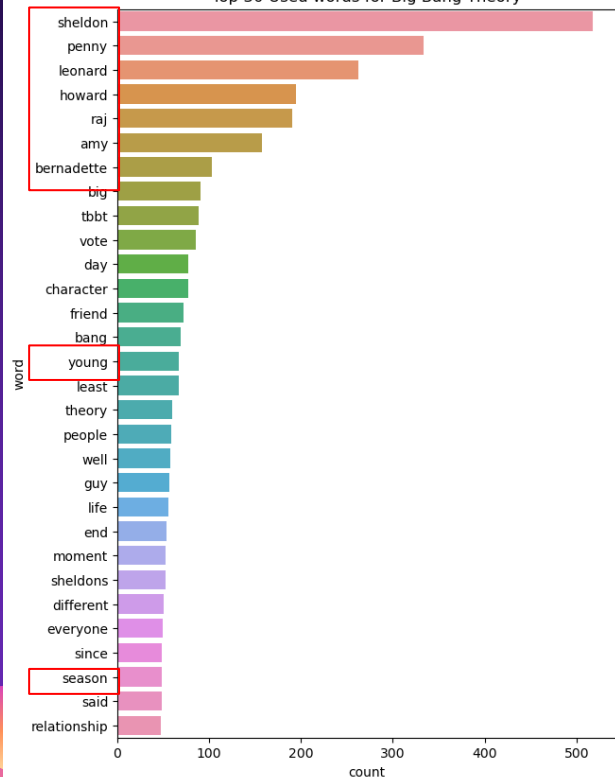


Common Words were observed for both subreddits

# Top 30 after Countvectorizer

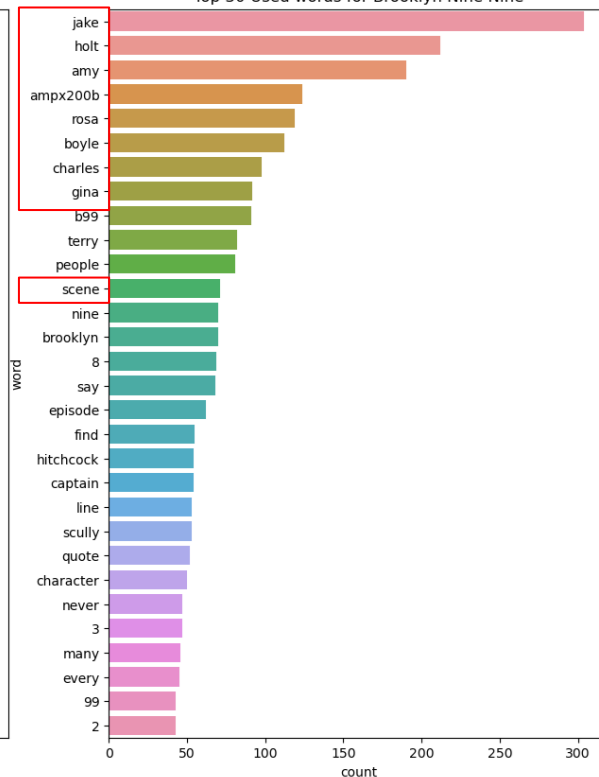
## r/bigbangtheory

Top 30 Used words for Big Bang Theory



## r/brooklynennine

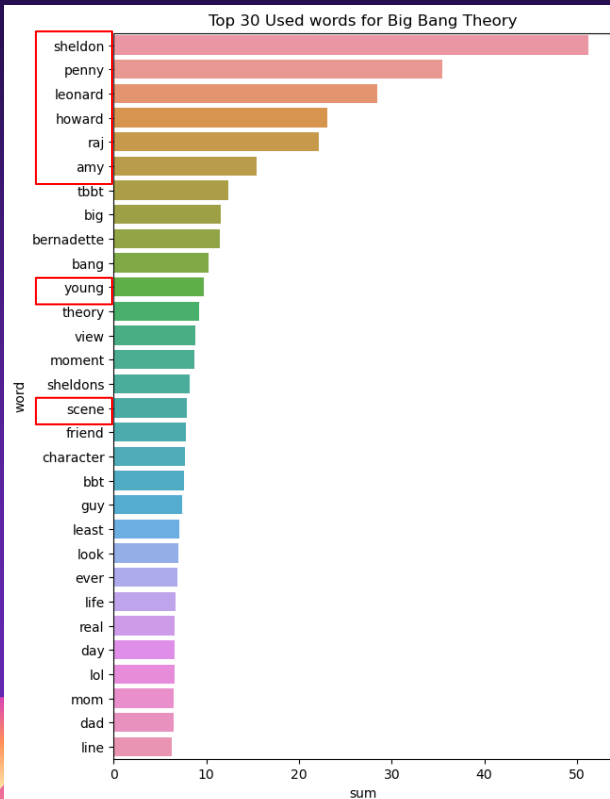
Top 30 Used words for Brooklyn Nine Nine



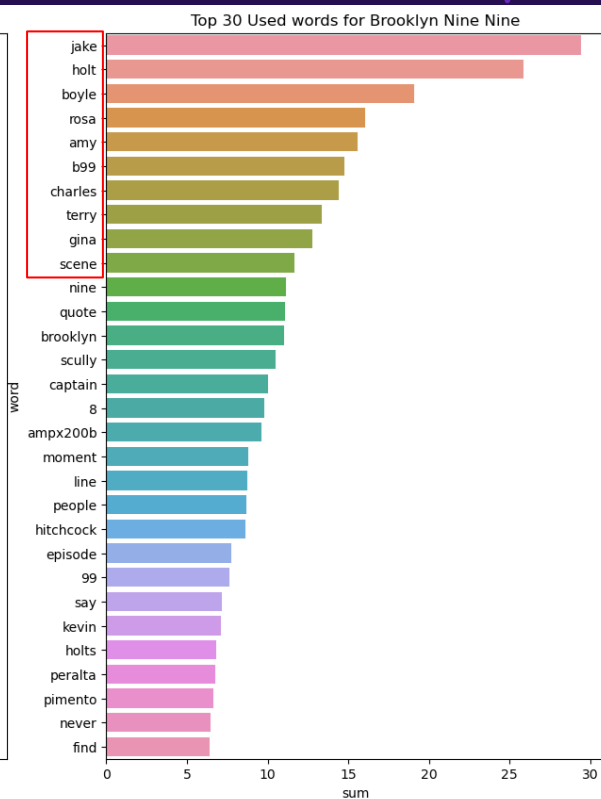
1. Character names are widely observed
1. 'Young' is seen on BBT very often due to sequel
1. 'Season' in BBT vs 'Scene' in B99

# Similar Observations are found after TF-IDF Vectorizer with the same stopwords

r/bigbangtheory



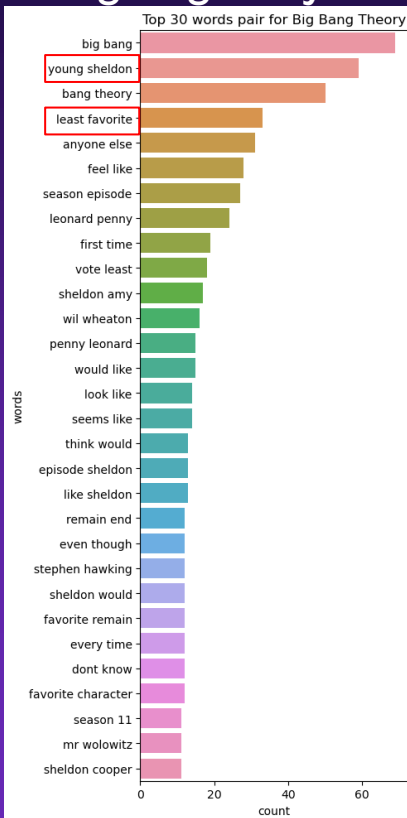
r/brooklynynine



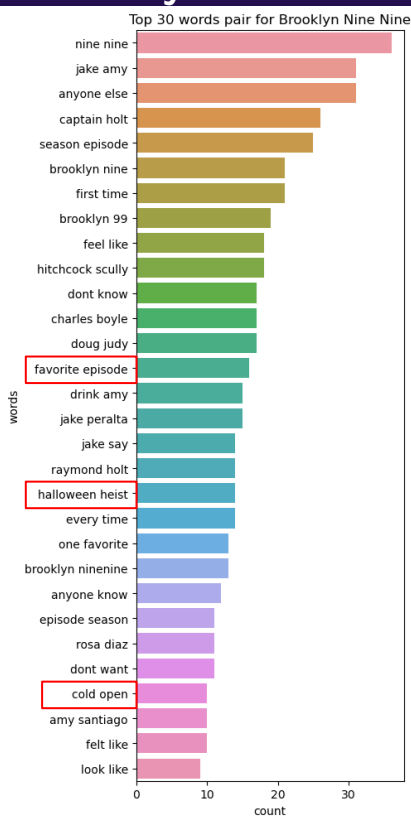
1. Character names are widely observed
1. 'Young' is seen on BBT very often due to sequel
1. 'Season' in BBT vs 'Scene' in B99

# Top 30 Common Words after Bigrams

r/bigbangtheory



r/brooklynyninenine



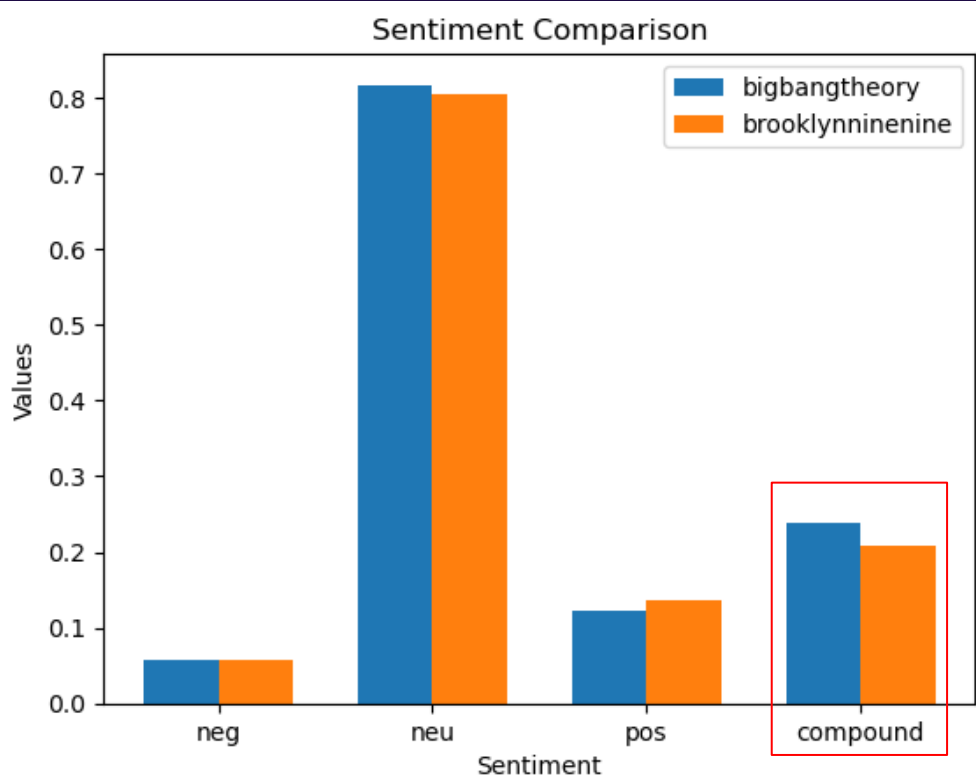
1. 'Least Favourite' in BBT vs 'favourite episode' in B99

1. 'Sheldon' has strong impact on viewers

1. 'Cold Open' is a unique X-factor in B99

1. 'Halloween Heist' is also a unique X-factor in B99

# Sentiment Comparison



1. Majority are **neutral** words

1. Positive are **2x** of Negative

1. BBT has **SLIGHTLY HIGHER** positive response than B99

‘Compound’ is **Aggregate Score** of the words

03

# Classification Model





“Disclaimer: This model is trained to predict the classification of a subreddit post based on the **fans’ comments** about the show, and not the **content of the show** itself.

Ming Fatt et-al, 2023

# What did we do?

18

## Models

18 Models were made during the process of training and building this classification model

>9k

## Hyperparameter tuning

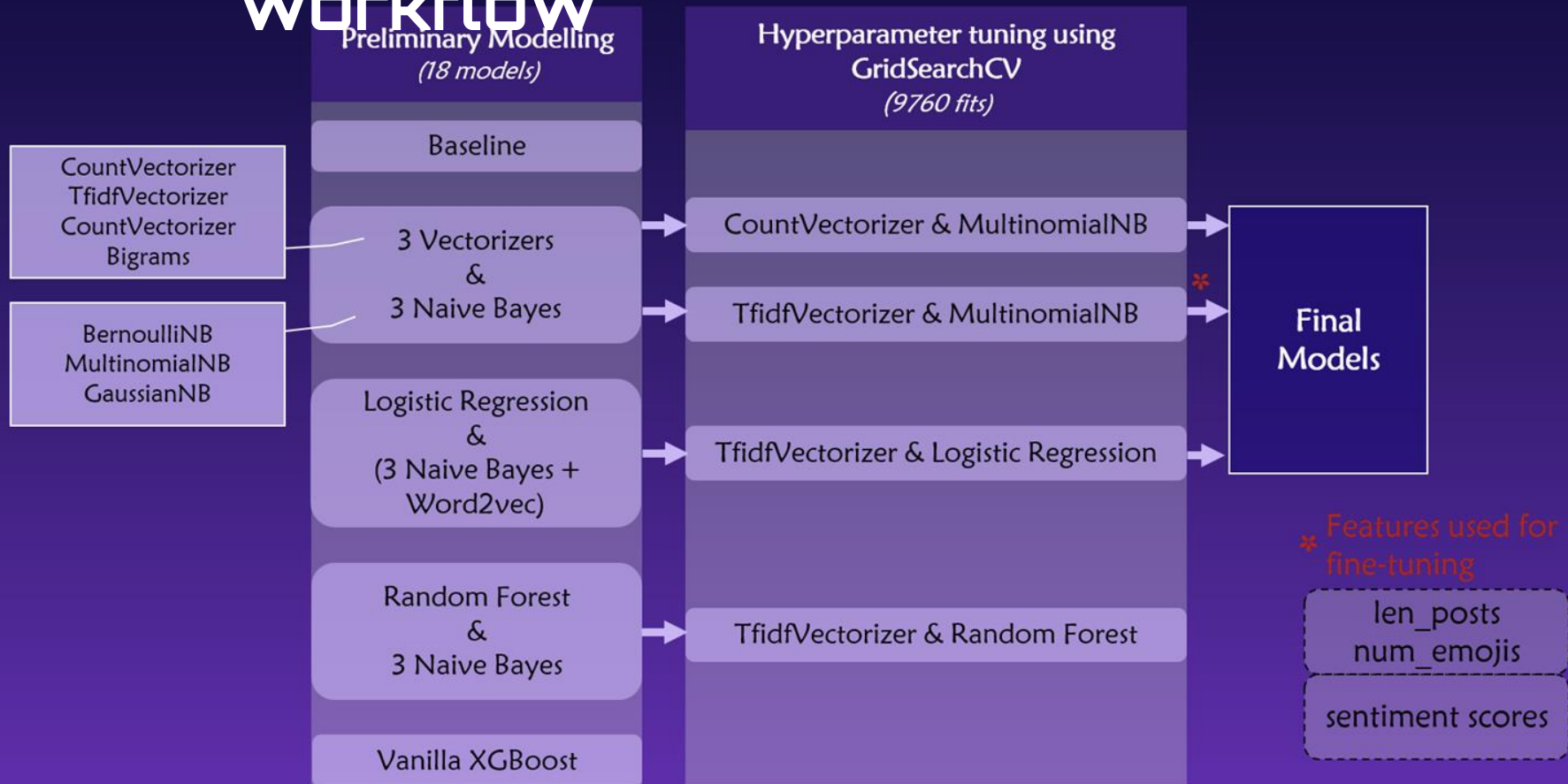
More than 9,000 fits was attempted to tune the hyper parameters of the model

01

## Baseline XG Boost

Boosting method was attempted to increase the performance of the model

# Modelling workflow



# Summary Table Score

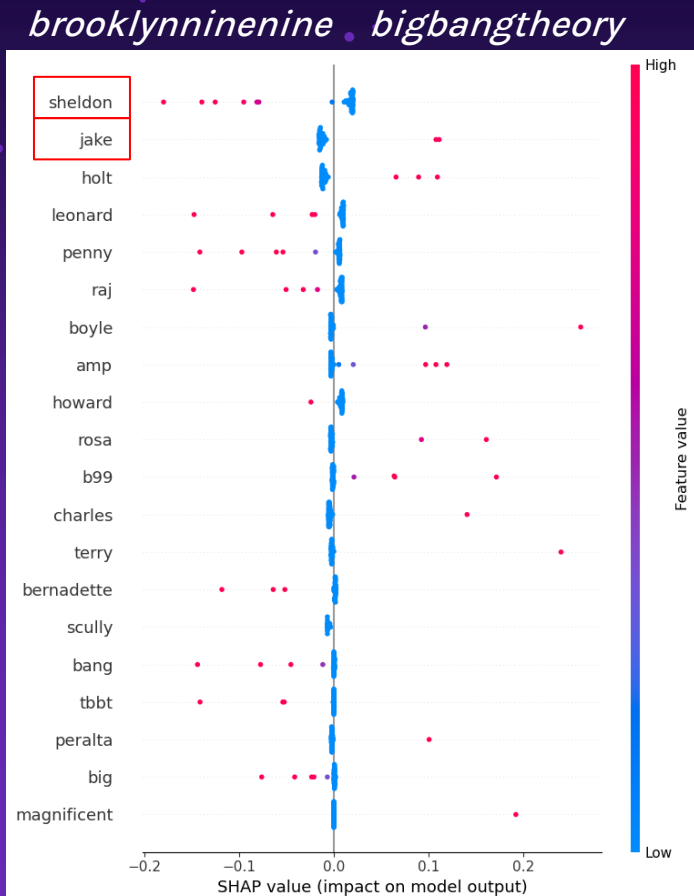
	Accuracy (Train)	Accuracy (Test)	Cross Validation Score
Baseline Model	0.520833		NA
Multinomial(NB) + CountVect + GridSearchCV	0.97050	0.90653	0.88123
Logistic Regression + TF-IDF + GridSearchCV	0.97957	0.90652	0.87368
Multinomial(NB) + TF-IDF + GridSearchCV	0.98865	0.92063	0.87973

*Best Model Hyperparameters:*

*TfidfVectorizer: max\_df=0.9, max\_features=max\_features, min\_df=1*

*MultinomialNB: alpha=0.5, fit\_prior=True*

# Features Importance



The character importances was doubly assured by the SHAP plot.

Sheldon is the most important feature for bigbagtheory

Jake is the most important feature for brooklynninenine

# Features Importance (Without Characters)

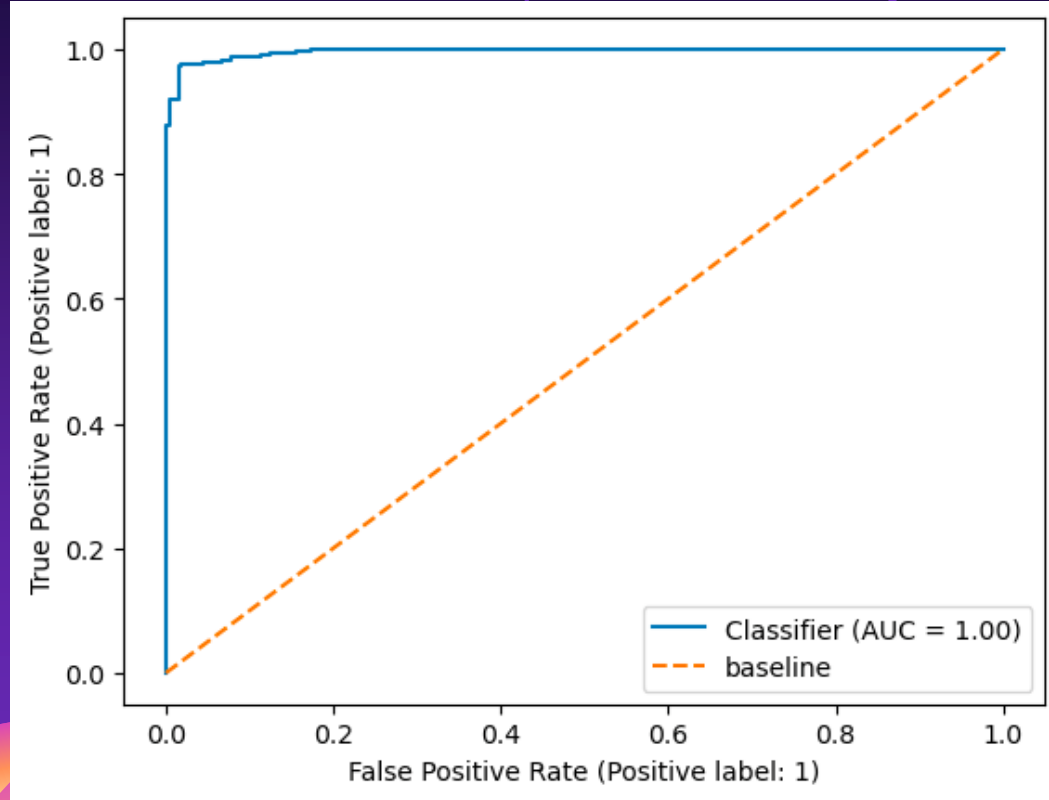
*brooklynninenine*   *bigbangtheory*

tbbt	-2.880121
view	-2.583689
bbt	-2.308415
bang	-2.134076
vote	-2.097841
gang	-2.031267
father	-1.924593
flag	-1.886893
comet	-1.801713
science	-1.709959
knock	-1.708414
young	-1.665038
nobel	-1.623368
hbo	-1.542135
dating	-1.538068
star	-1.521292
max	-1.444057
big	-1.434441
coffee	-1.429333
prize	-1.396606

b99	3.041289
nine	2.887422
brooklyn	2.878999
99	2.365311
captain	2.128342
ninenine	2.067835
heist	1.933225
cop	1.774538
magnificent	1.747412
police	1.705123
blank	1.683344
8	1.675029
detective	1.668509
fill	1.651269
tonight	1.628251
quote	1.627586
precinct	1.615997
drink	1.568819
cheddar	1.559990
prison	1.550717

1. There are distinct features from each class.
1. Positive: Class 1  
Negative: Class 0
1. HBO is mentioned but netflix is not.

# ROC AUC curve



1. The bigger the area under curve, the better our model accuracy.
1. The best score one can get in AUC Classifier is 1.
1. Our best model achieved an accuracy of 0.996, rounded to 1 by machine.

# Confusion Matrix

		Actual	
		Positive (1)	Negative (0)
Predicted	Positive (1)	TP=268	FP=4
	Negative (0)	FN=8	TN=287

## Misclassified posts analysis

- For FP and FN posts, after TF-IDF Vectorisation was carried out, only the word 'season' was wrongly classified
- 'Season' is in top 30 common words for both sitcoms



04

# Product Demonstration



05

## Recommendation and Key Insights



## Project Goals Revisit:

- a. Identify what show elements in the sitcom are popular among the viewers
- a. To build an infrastructure that can help to classify and analyse user's comments about the show from various platforms.

# Key Insights

## Show Elements

'Cold Open'  
'Halloween Heist'  
'Potential Sequel'

## Hot Topics

Favorite Character  
Least Favorite Scene

## Character Names

All characters are  
repeatedly mentioned

# Key Recommendations

In new shows/future seasons,

01

Build Memorable  
Characters

02

‘Cold Open’ for  
every episode

03

Periodical events like  
‘Halloween Heist’

The characters come with a moment/scene!

## Future Work

- Model can be expanded to Multi-Class Classification
- Further collect text inputs from other sources periodically
- To perform further sentiment analysis

06

# Key Limitations





# Misclassified Words

Predicted as B99 with 51.51% probability.

Caltech

Supersymmetry

Anakin Skywalker

Yoda

Amy

Cheese Cake



## Future Work

- Model can be expanded to Multi-Class Classification
- Further collect text inputs from other sources periodically
- To perform further sentiment analysis



Thank you!



# END

Team Data-Nine-Nine:

Ming Fatt

Jin Jin

Jasmine

Willson

Wenxi