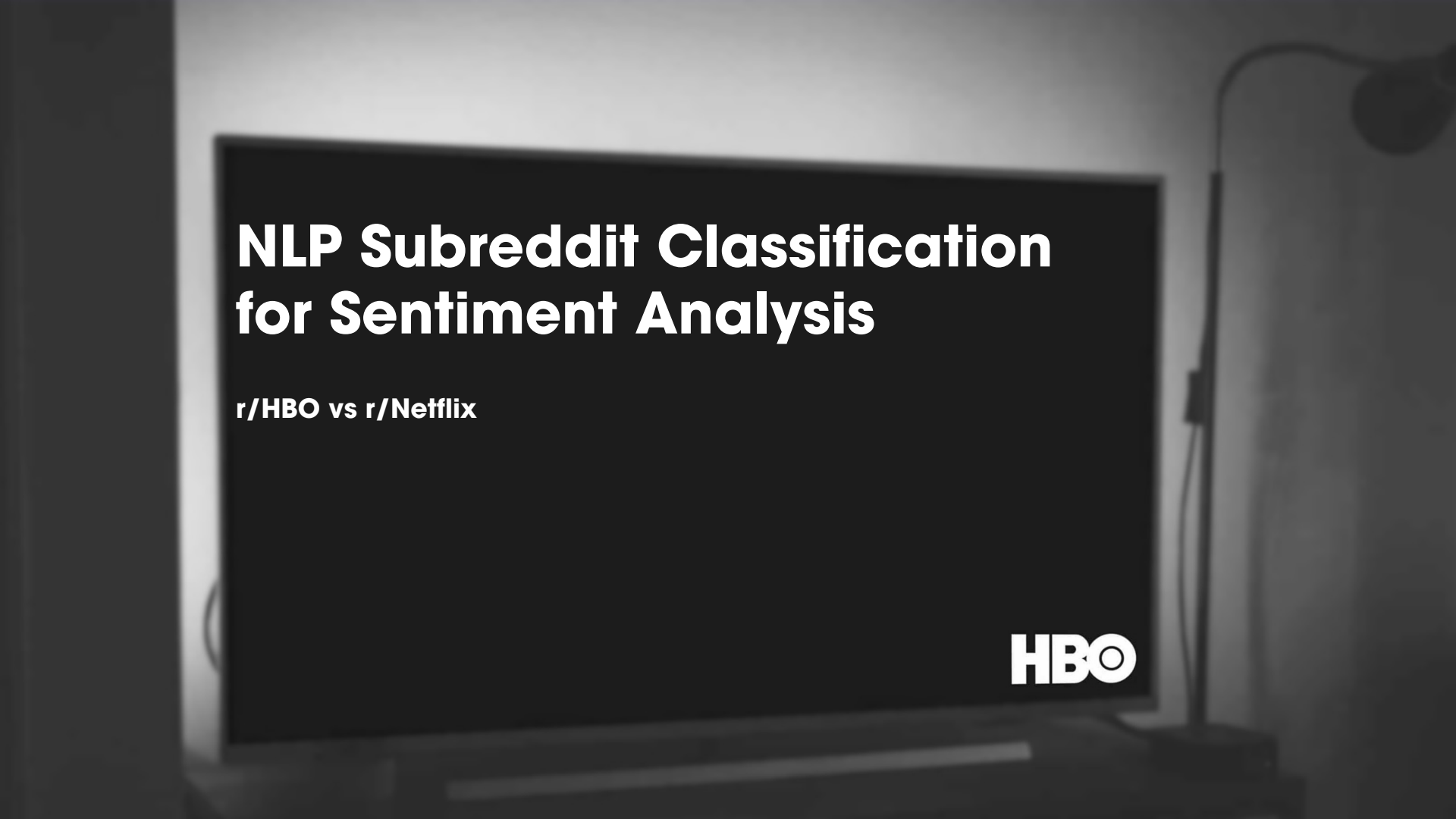
NLP Subreddit Classification for Sentiment Analysis

r/HBO vs r/Netflix



Agenda

01 Background

02 Problem Statement

03 Workflow

04 Data Cleaning

05 Exploratory Data Analysis

06 Modeling

07 Limitations

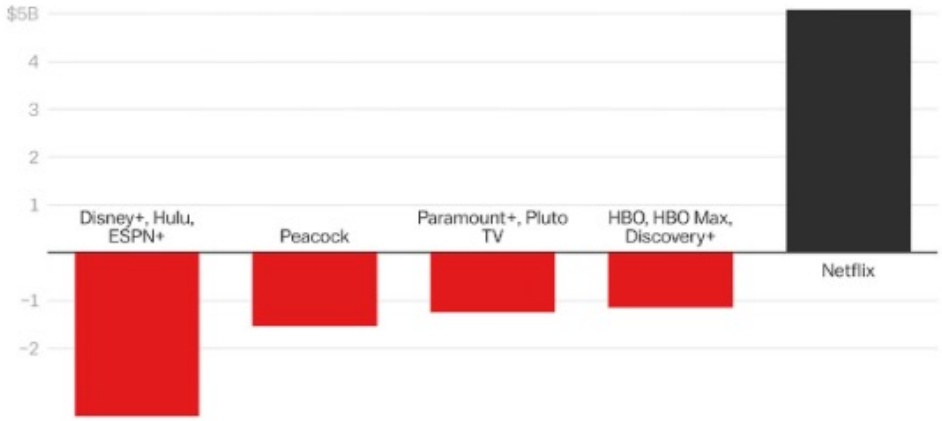
08 Recommendations & Conclusion

Background

HBO VS NETFLIX

Netflix competitors are losing billions trying to catch the streaming giant

Profit/loss for the first three quarters of 2022 (HBO totals are for quarters two and three)



Disney and Netflix data is operating loss/income; Paramount uses OIBDA; Peacock and HBO report adjusted EBITDA.

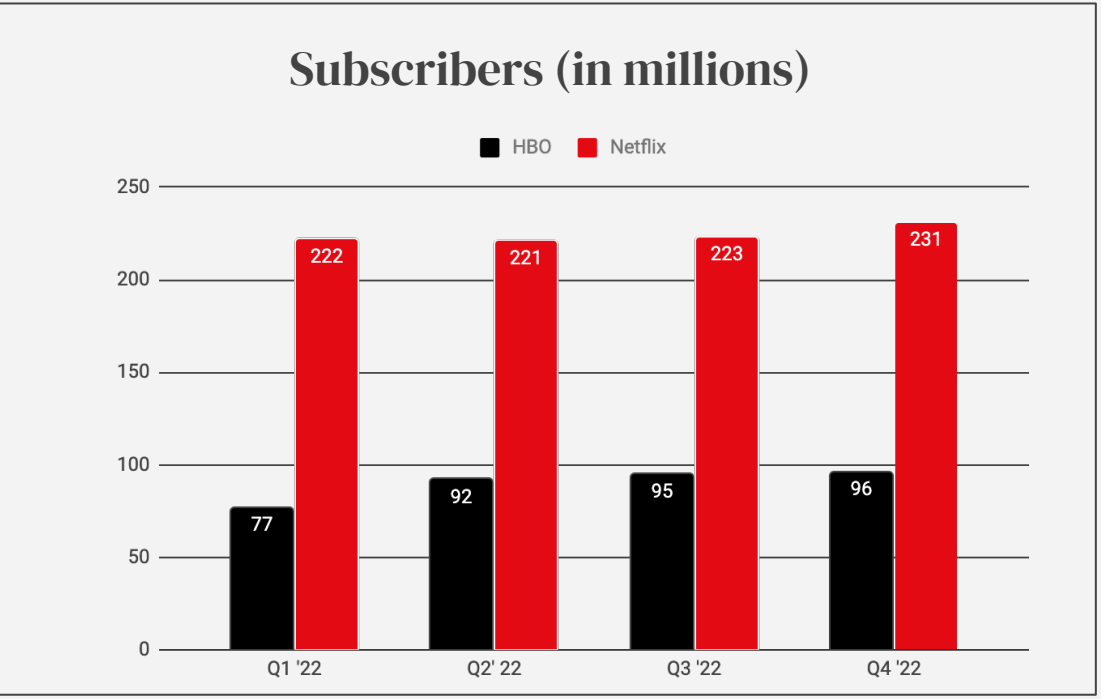
Source: Company filings

recode BY Vox

HBO has lost billions of dollars in an effort to play catch up to its main rival

Background

HBO VS NETFLIX



Why is this happening?

- Over-reliance on its traditional cable TV model
- Significant investments in developing streaming platform and producing original content

What can be done?

Audience Targeting

Gain insights on viewer preferences and tailor marketing campaigns accordingly

Content Development

Identify emerging trends and topics

Competitor Analysis

Identify areas of differentiation and competitive advantage

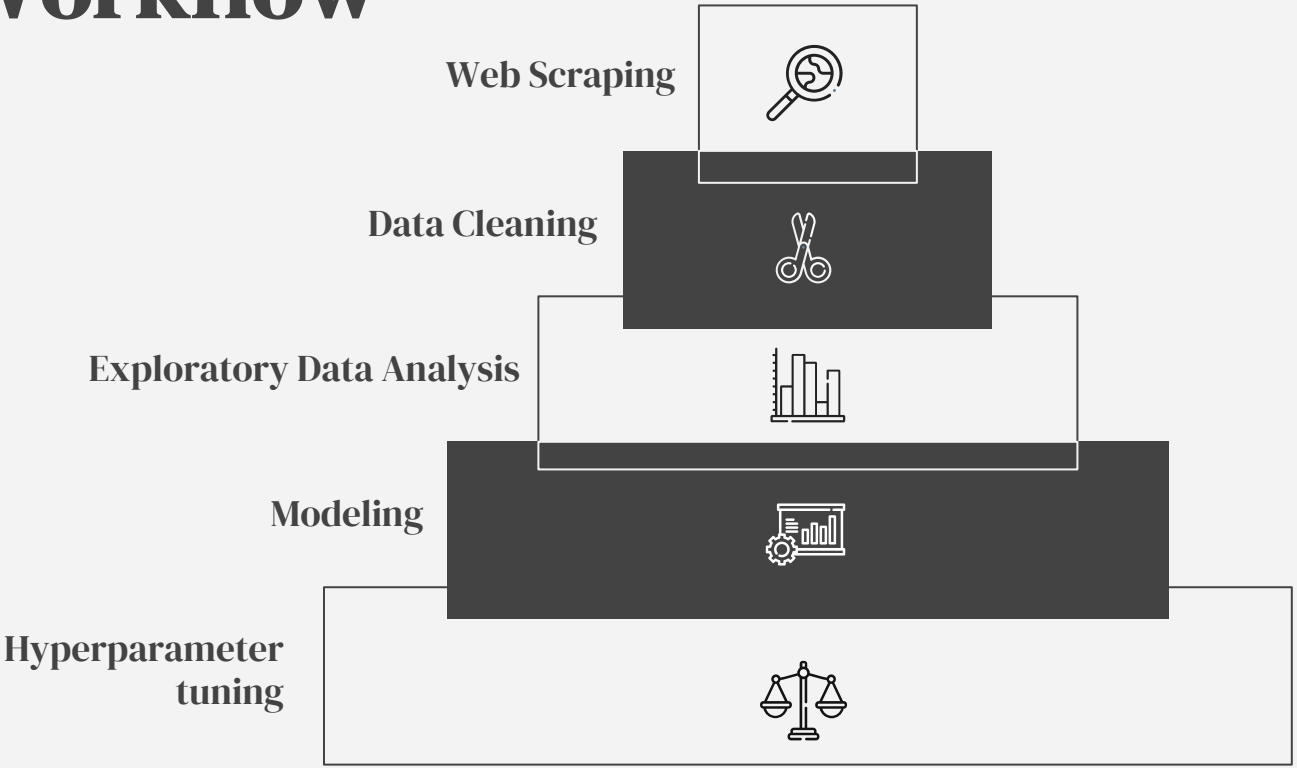
Sentiment Analysis

Enhance its offerings and strengthen relationship with audience

Problem Statement

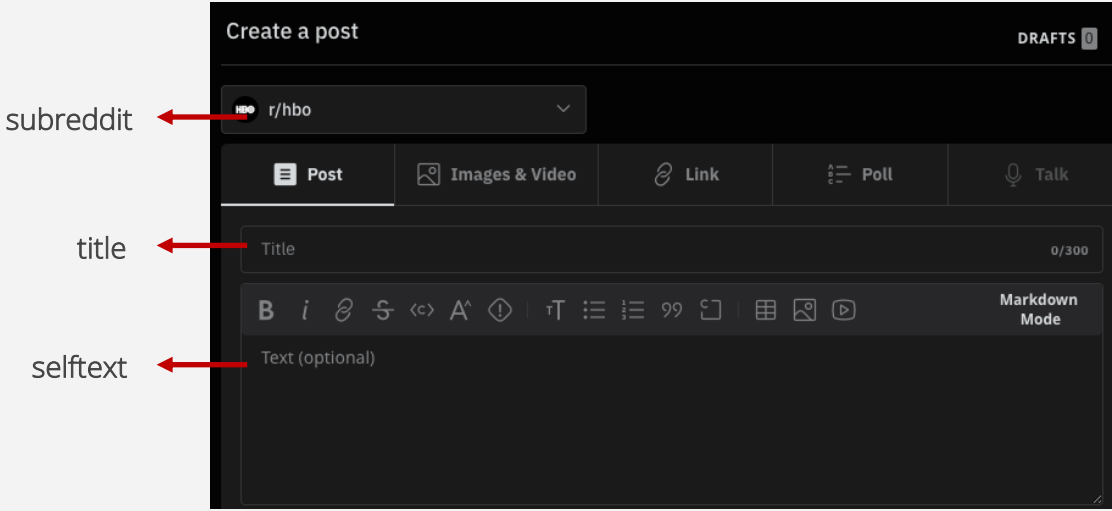
As part of the Data team at HBO, this project aims to build a machine learning model to classify subreddits based on their content to identify topics that generate the most engagement and discussion among viewers. These insights will help the company make more informed decisions about its content development, marketing, and overall strategy.

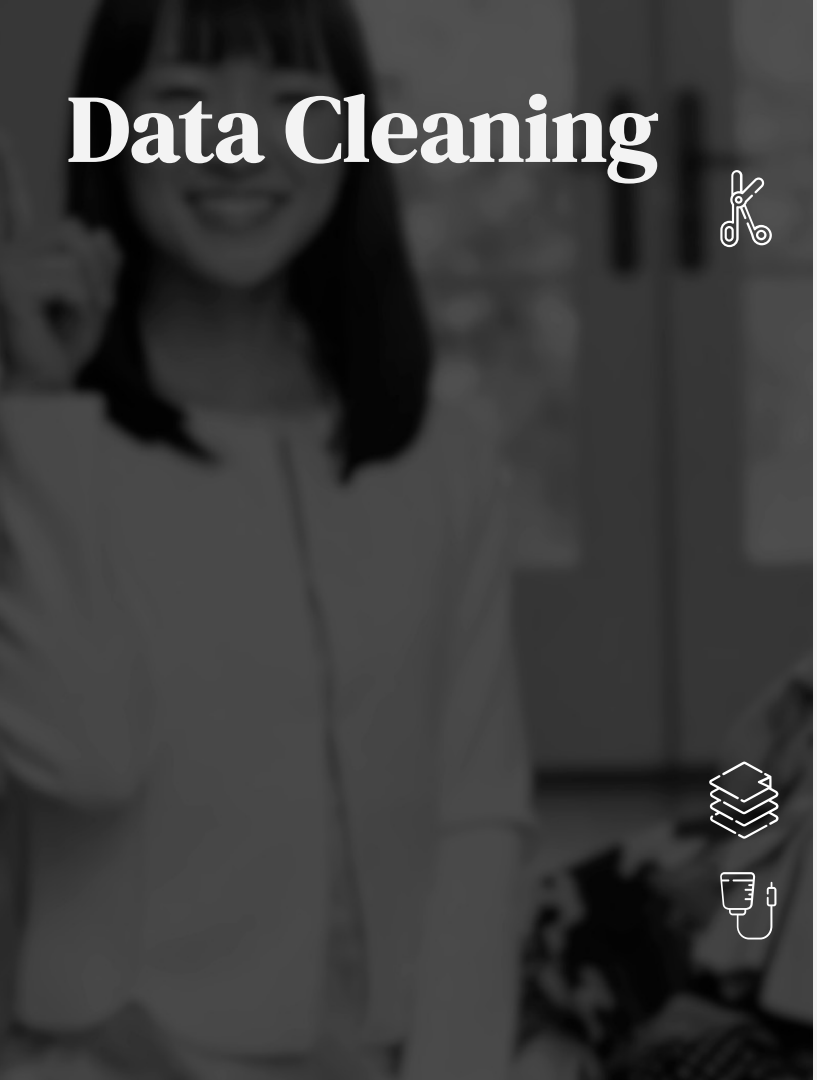
Workflow


Web Scraping
Data Cleaning
Exploratory Data Analysis
Modeling
Hyperparameter tuning

Web Scraping using Pushshift API

- 1,000 posts retrieved per subreddit
- 3 columns





Data Cleaning

Remove the following:

- Null / duplicate values
- [removed] and [deleted] values
- Punctuation
- HTML text
- Non-English text
- Non-alphanumeric characters
- Stopwords

Tokenisation

Lemmatisation

Examples
removing...

Non-alphanumeric chars

I stayed up all night making this. 💔

I stayed up all night making this

HTML text

My top 10 HBO series 1. The Wire\n2. Chernobyl...

My top 10 HBO series 1. The Wire 2. Chernobyl ...

[removed] values

[removed]

Punctuation

[Selling] Netflix UHD month for $

Selling Netflix UHD month for

Duplicates

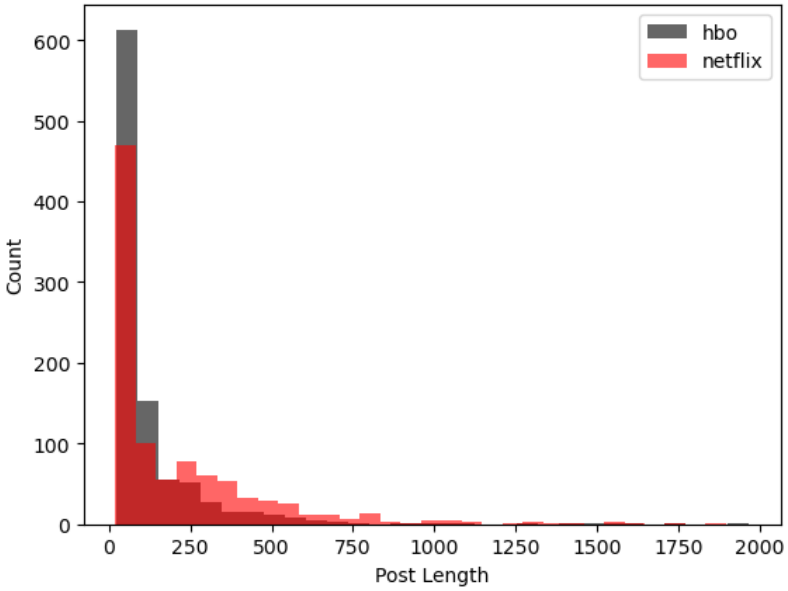
Is it me or does netflix try very heard to fin...

Is it me or does netflix try very heard to fin...

Tokenisation & Lemmatisation

Exploratory Data Analysis

- 9 posts greater than 2,000 were removed as it was deemed irrelevant and could potentially skew our model
- Netflix has greater post lengths overall



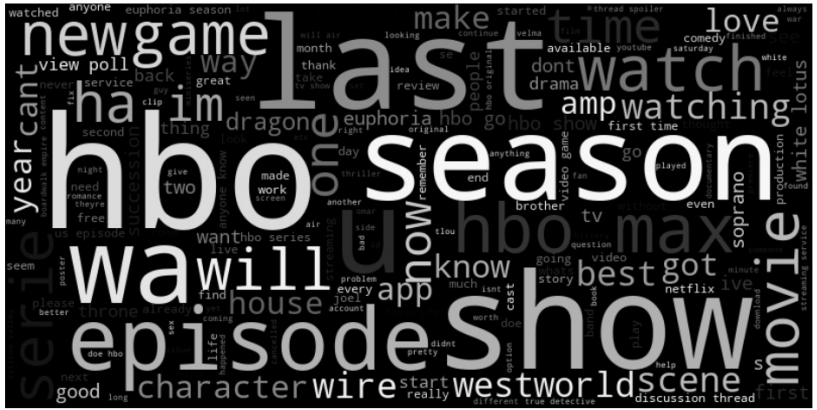
HBO	
count	975.0
mean	139.7
std	208.9
min	24.0
25%	50.0
50%	71.0
75%	129.0
max	1968.0

Netflix	
count	981.0
mean	213.9
std	261.9
min	20.0
25%	50.0
50%	89.0
75%	293.0
max	1897.0

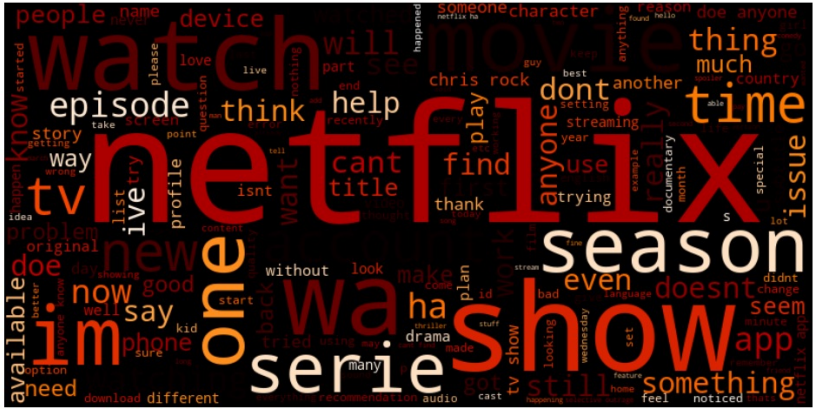
Exploratory Data Analysis

- WordClouds were utilised to check for commonly occurring words to add to our Stopwords
- For example, ‘view poll’ was dropped as these are common among “poll” Reddit posts

HBO Word Cloud



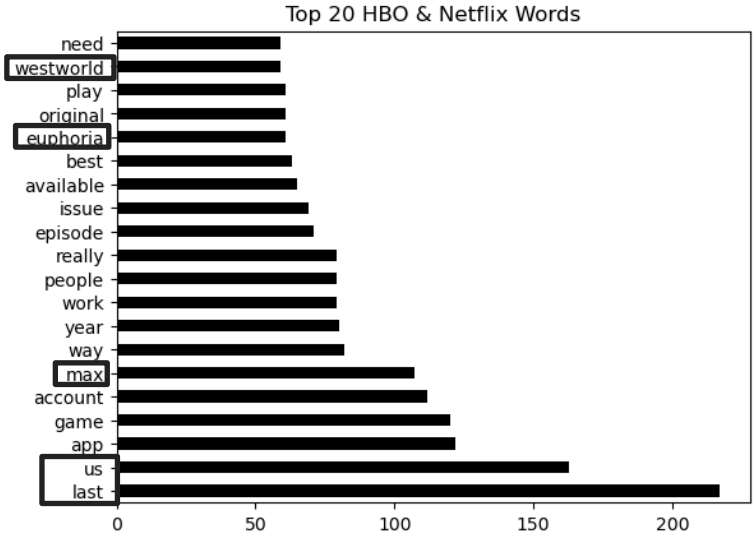
Netflix Word Cloud

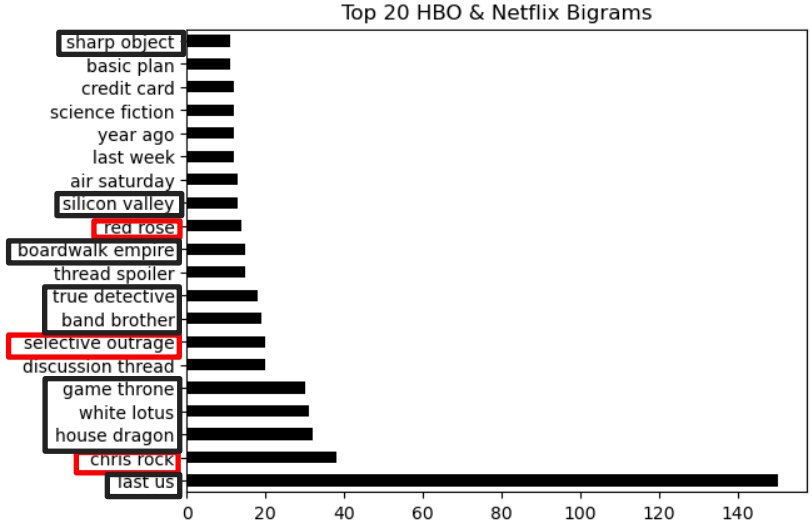


Exploratory Data Analysis

HBO Netflix

- Interestingly, most of the keywords spotted relate to HBO
- There seems to be many non-content related keywords that appear here
- HBO has 11 shows mentioned that rivals Netflix's 3

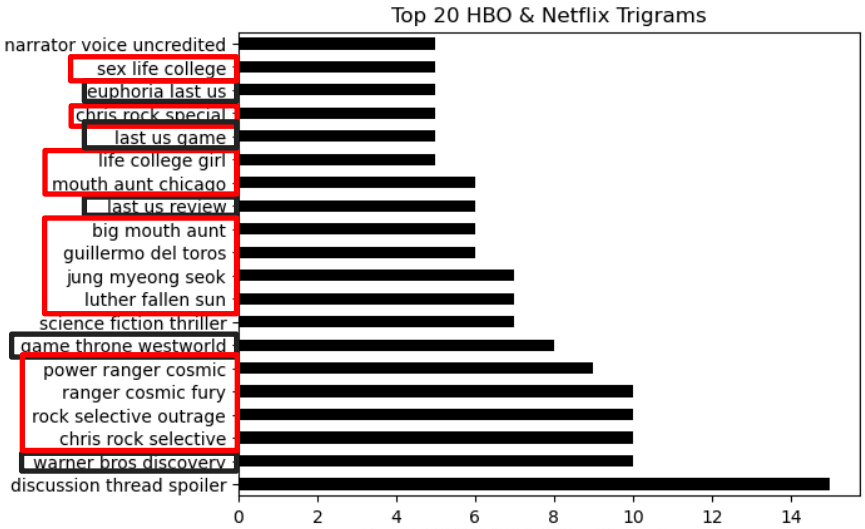


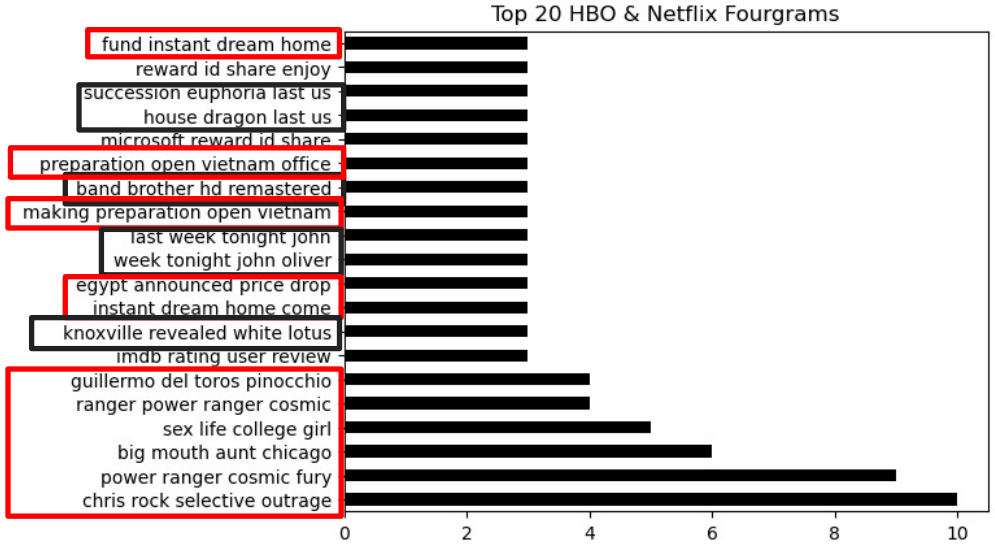


Exploratory Data Analysis

HBO Netflix

- Netflix has 9 show mentions to HBO's 8
- However, number of keyword occurrences here are not as significant





Modeling

Preprocessing / Data Cleaning → **Classification Model Selection** → **Model Optimisation**

- CountVectorizer
- N-grams (baseline only)
- TF-IDF Vectorizer

- Multinomial Naïve Bayes
- Logistic Regression
- Random Forest
- Gradient Boosting
- Linear SVC

- Hyperparameter Tuning

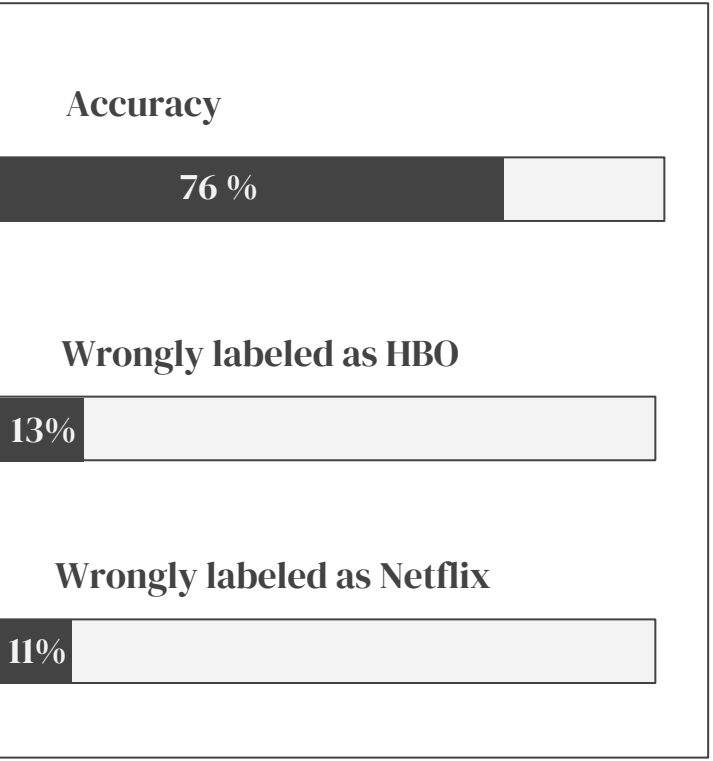
Modeling – Overview

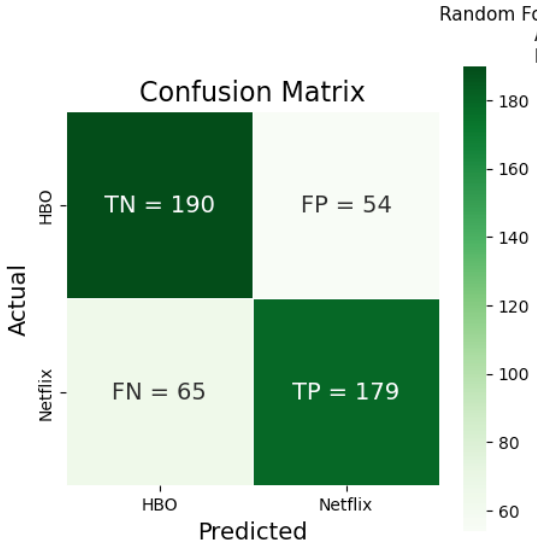
- Multinomial Naïve Bayes
- Logistic Regression
- Random Forest
- Gradient Boosting
- Linear SVC

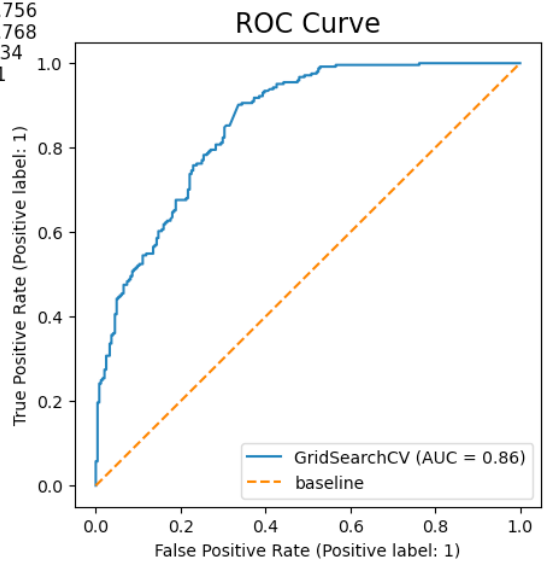
Model Results (in %)

	Before Tuning		After Tuning	
Vector / Model	**Train Accuracy**	**Test Accuracy**	**Train Accuracy**	**Test Accuracy**
Multinomial Naïve Bayes CV	94.8	78.4	99.6	80.3
Multinomial Naïve Bayes TF-IDF	95.7	80.2	99.9	81.8
Logistic Regression CV	99.3	77.6	99.0	79.9
Logistic Regression TF-IDF	96.4	78.0	99.9	81.1
Random Forest CV	99.3	75.7	84.9	75.6
Random Forest TF-IDF	99.9	76.9	83.1	73.2
Gradient Boosting CV	81.5	75.1	89.5	76.0
Gradient Boosting TF-IDF	83.2	75.5	89.1	73.4
Linear SVC CV	99.9	75.1	92.0	77.0
Linear SVC TF-IDF	99.8	78.0	99.9	80.9

76% of predictions are correctly labeled







Limitations

Limited training data

More data is required to train the model (eg. lingo/ acronyms)

Limited resources

May require significant computing power with larger datasets

Sentiment Analysis

Model requires more features to better gauge sentiment

Inherent Variability

Variability of random forest model can lead to different results each time model is run

Recommendations

- The Last of Us, House of the Dragon and White Lotus were the most popular shows among HBO fans in the past year
- Further analysis using NLP techniques should identify which aspects of shows resonate most with fans
- Inform content development decisions, marketing messaging and product differentiation
- HBO's consistent YoY increase in revenue, subscriber numbers and market share is a healthy sign that it should continue to focus on quality over quantity to cater to its ardent fanbase

Thank you

HBO

All images taken from HBO