Streaming Wars

Subreddit Analysis

Irvan Indra Wahab

Disney+ Team

The Walt Disney Company

Agenda

01
Introduction
Problem statement, overview, and objectives

02
Methodology
Data science process and procedure

03
Data Preparation
Data collection, cleaning and preprocessing

04
EDA
Data exploration and visualization

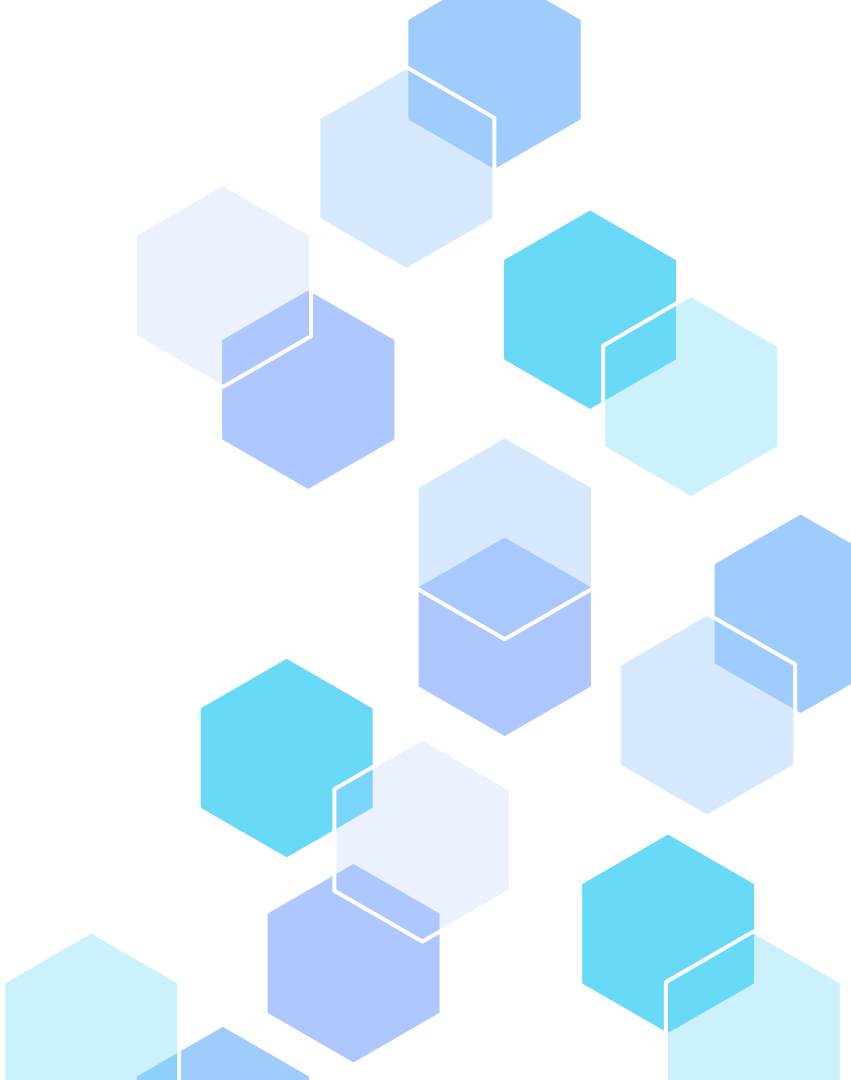
05
Modeling
Model selection, training, and hyperparameter tuning

06
Evaluation
Model performance evaluation and recommendation

01

Introduction

Problem statement, overview, and objectives



Problem Statement

Overview

Disney+ was launched in Nov-2019, and has quickly become a major player in the streaming industry

Problem

Formidable challenge to effectively compete with Netflix's market dominance

Objective

Gain understanding of public perception, user preferences, and emerging trends through Reddit

149.6 Million+
Subscribers (*As of December 2023)

13,000+
Shows & Movies

8,000
Hours of Content

150+
Markets

39
Languages

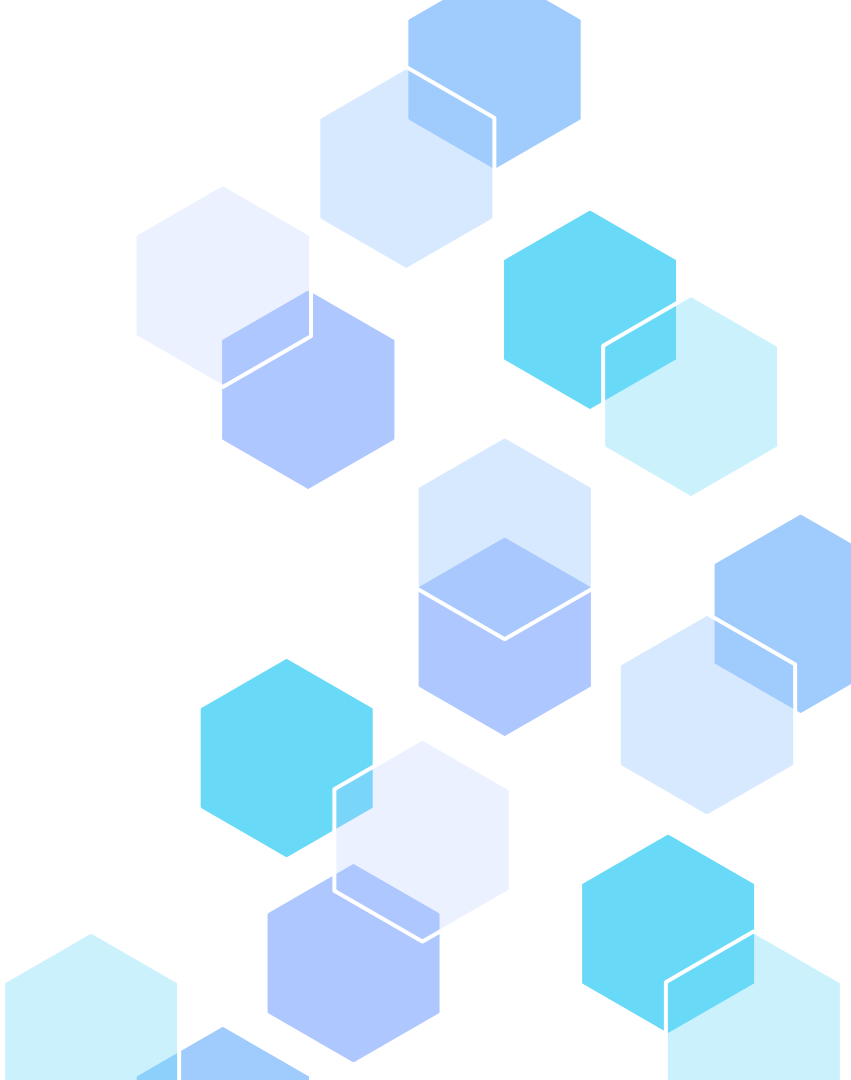
Product Comparison

Disney+	vs	Netflix
149.6 Million	**Subscribers**	260.3 Million
$ 7.0 monthly	**Subscription Fee**	$ 8.0 monthly
$ 8.4 Billion	**2023 Revenue**	$ 33.7 Billion

02

Methodology

Data science process and procedure



Methodology

Collection

Web scraping of Subreddit posts

Cleaning

Remove duplicates and empty row

Preprocessing

Feature and variables selection

Exploratory

Data exploration and visualization

Modeling

Model selection, fitting and training

Evaluation

Model performance assessment

03

Data Preparation

Data collection, cleaning and preprocessing

Public perception from Reddit

Every day, millions of people around the world post, vote, and comment in communities organized around their interests.

Reddit by the Numbers

Reddit is a growing family of millions of diverse people sharing the things they care about most.

As of December 31, 2023

73M+	267M+	100K+	16B+
Daily Active Uniques	Weekly Active Uniques	Active Communities	Posts & Comments

Post

The community can share content by posting stories, links, images, and videos.

Comment

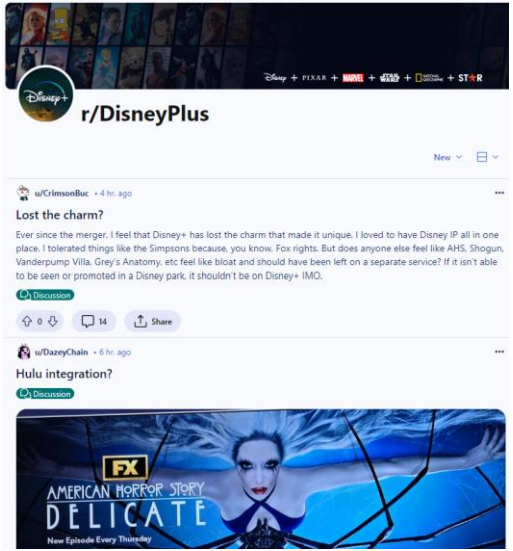
The community comments on posts. Comments provide discussion and often humor.

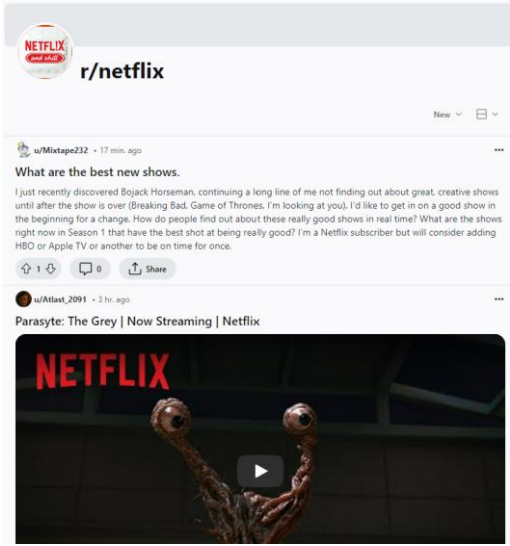
Vote

Comments & posts can be upvoted or downvoted. The most interesting content rises to the top.

Data Collection - Subreddit

- Subreddits are user-created areas of interest where discussions on Reddit are organized, and are denoted by “r/”
- r/DisneyPlus vs r/netflix
- Scrape the content





Data Collection - Web Scraping

Title

Post title

Type

4 different URLs to scrape based on type, 250 posts each

Upvote Ratio

Ratio of upvotes to the total votes cast

Selftext

Body text of the post

Data Preparation

Data Cleaning

- Remove duplicate entries based on the user id post
- Remove the field with empty input

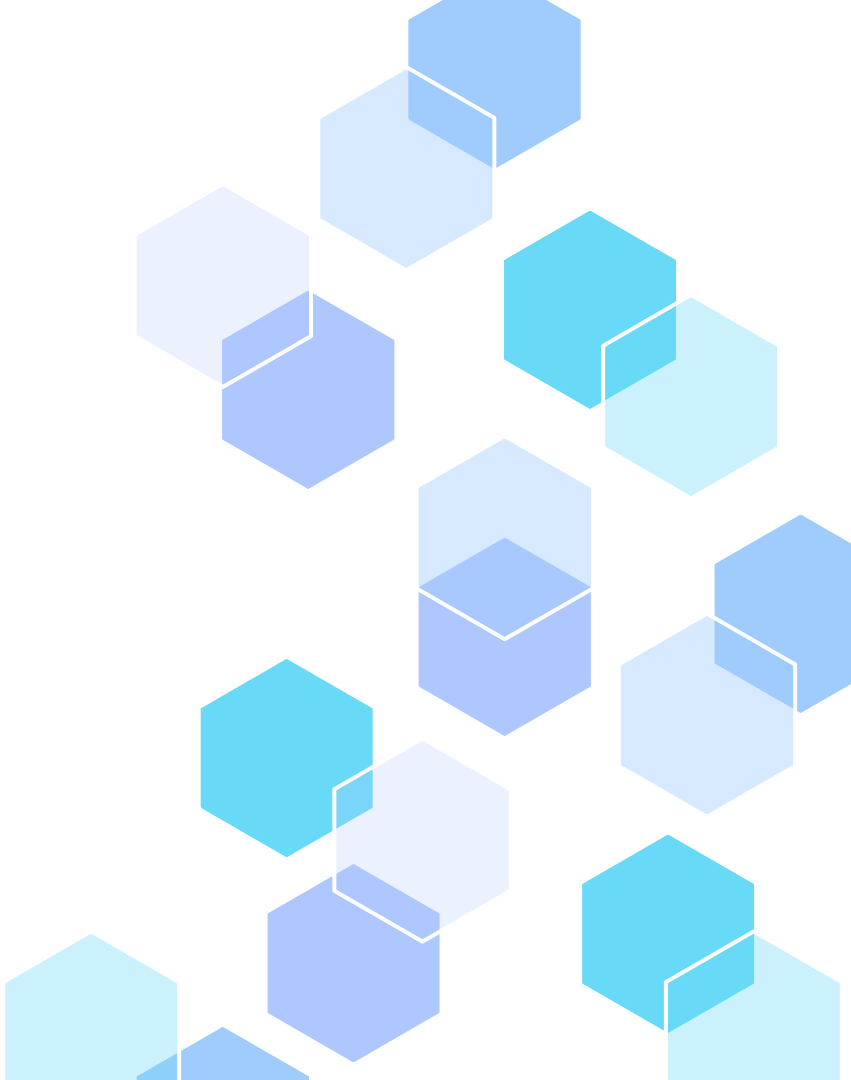
Preprocessing

- Combine Title and Selftext into one feature called “post” as our X variable
- Add feature called “label” to classify post as Y variable

04

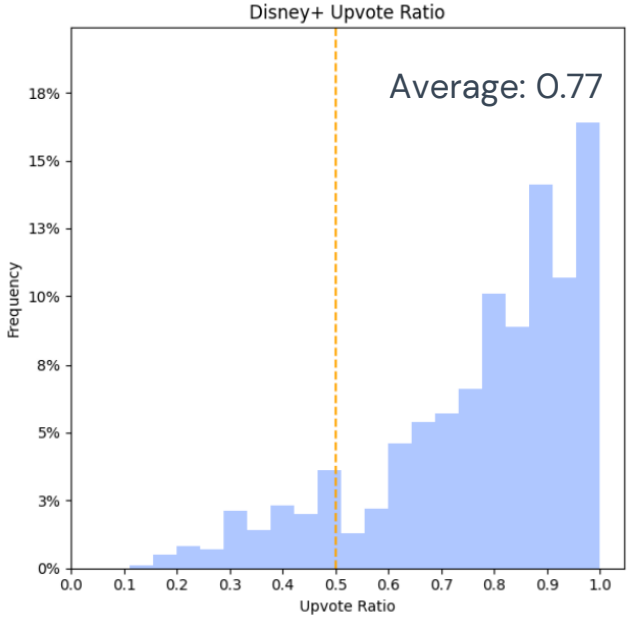
EDA

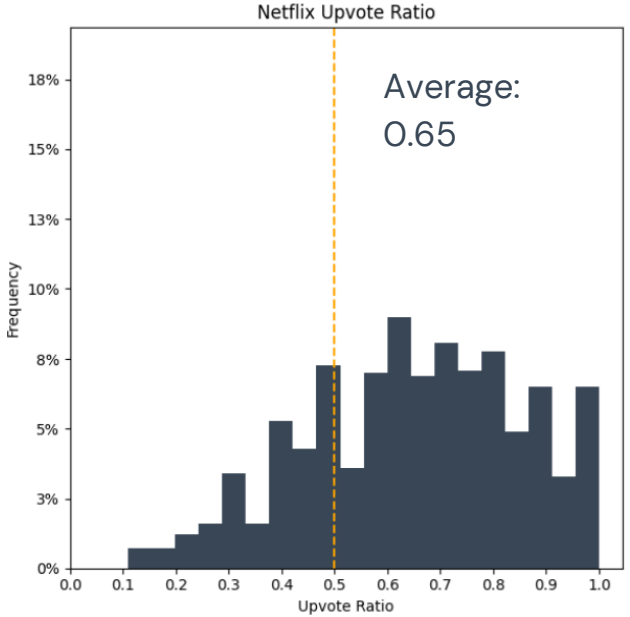
Data exploration and visualization



Data Exploration - Upvote Ratio

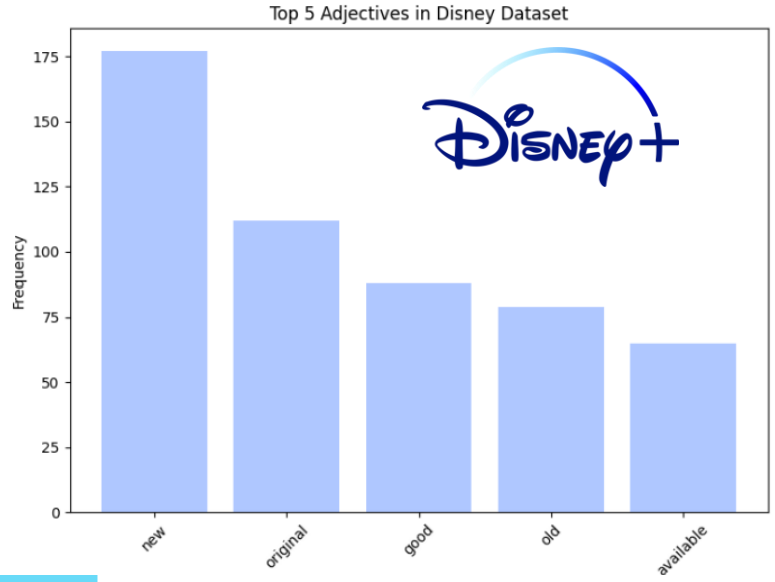
- Represents the ratio of upvotes to the total votes cast
- Majority of the posts have more upvotes - well liked by viewers

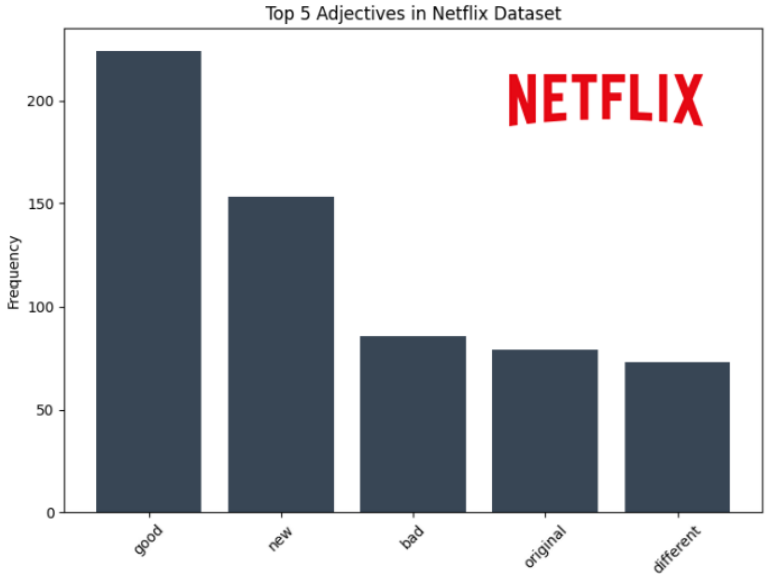




Data Exploration - Frequent Words

- Tokenize, lemmatize and remove stop words
- Disney+ has more “New” and “Original” but less of “Good”





Sentiment Analysis

0.187 — 0.204

Positive

Proportion of the text that expresses positive sentiment

0.083 — 0.124

Negative

Proportion of the text that expresses negative sentiment

1.0 — 1.0

Compound

Overall sentiment of the text, normalized between -1 and +1

05

Modeling

Model selection, training, and hyperparameter tuning

Modeling Overview

Train - Test

Split the data into train and test with equal proportion of Disney+ and Netflix on both sets (~50-50)

Model Fit

Fit different combinations of vectorizer and classifier

Tuning

Hyperparameter tuning on the best performing model combination

Modeling - Vectorizer

Algorithm used to convert text data into numerical vectors

CountVectorizer

Counting how many times each word appears and put into a table

TF-IDF

Assigning weights to each word based on its frequency

Modeling - Classifier

Algorithm trained to assign labels / categories to input data

Multinomial Naive Bayes

Logistic Regression

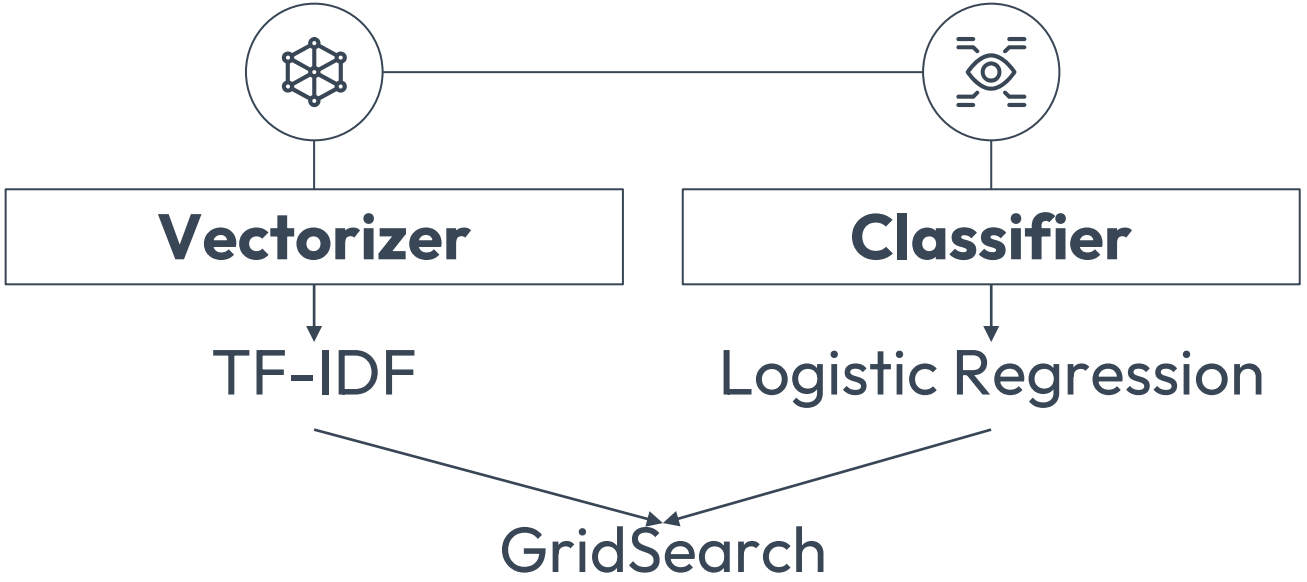
K-Nearest Neighbors

Random Forest

Bootstrap Aggregating

Gradient Boosting

Final Model

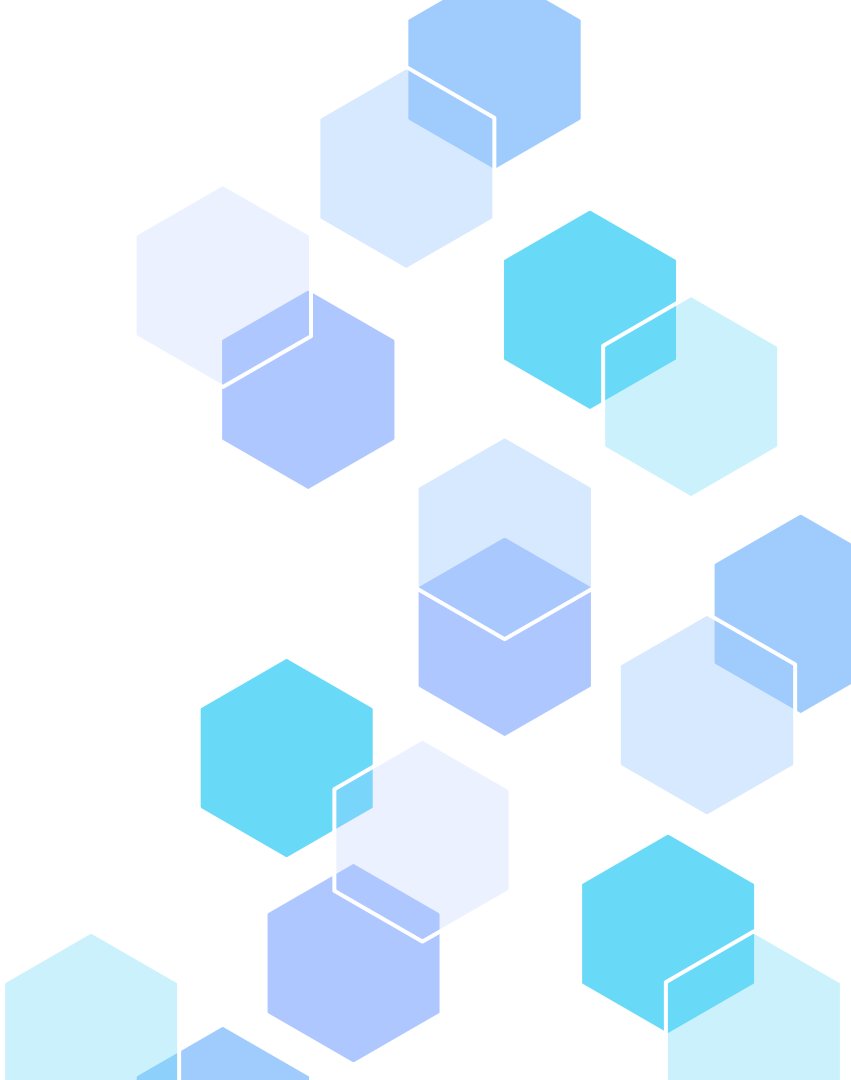


	Train	Test	Cross Validation
Accuracy	0.96	0.86	0.85

06

Evaluation

Model performance evaluation and recommendation



Evaluation - Metric

Accuracy

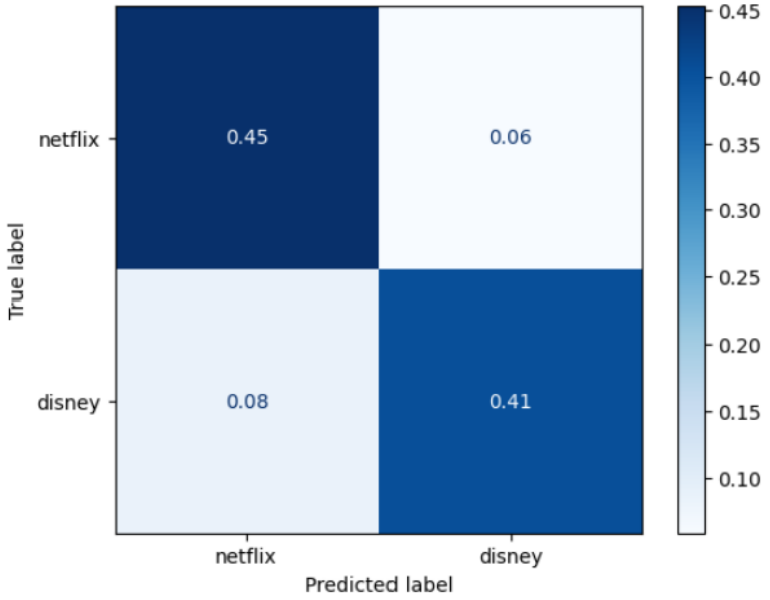
Test: 0.86

Proportion of correctly classified instances

Rationale

- Easy to interpret
- Good for balanced datasets
- Treats all prediction errors equally

Confusion Matrix



Recommendation

Data insights

Provide insights on public perception and preferences towards Disney+

Predictive modeling

Predict public opinion or comments and perform classification

Optimization

Optimize marketing and content creation strategy towards a more targeted public

Competition Analysis

Study the competitor dynamics, customer experience and preference

Limitations

Data Generation

Data scraped for this analysis are <2,000 - to scrape more data on further studies / analysis

Variability

Reddit users come from diverse backgrounds and have different type of writing styles

Biases

Potential biases amongst communities towards certain items

Thank you!