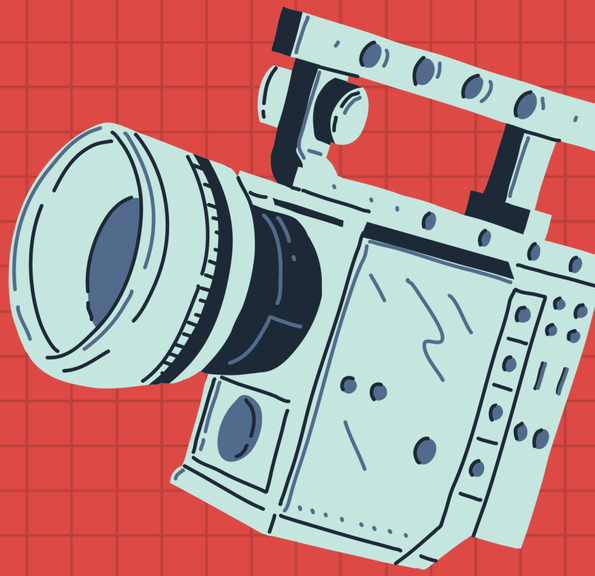
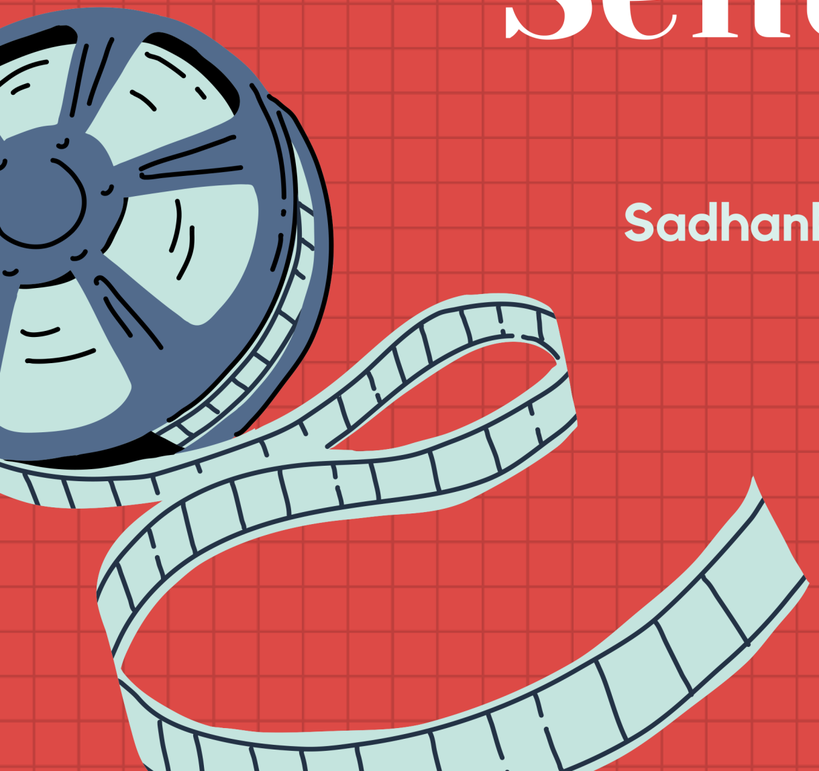


Movie Review Sentiment Analysis



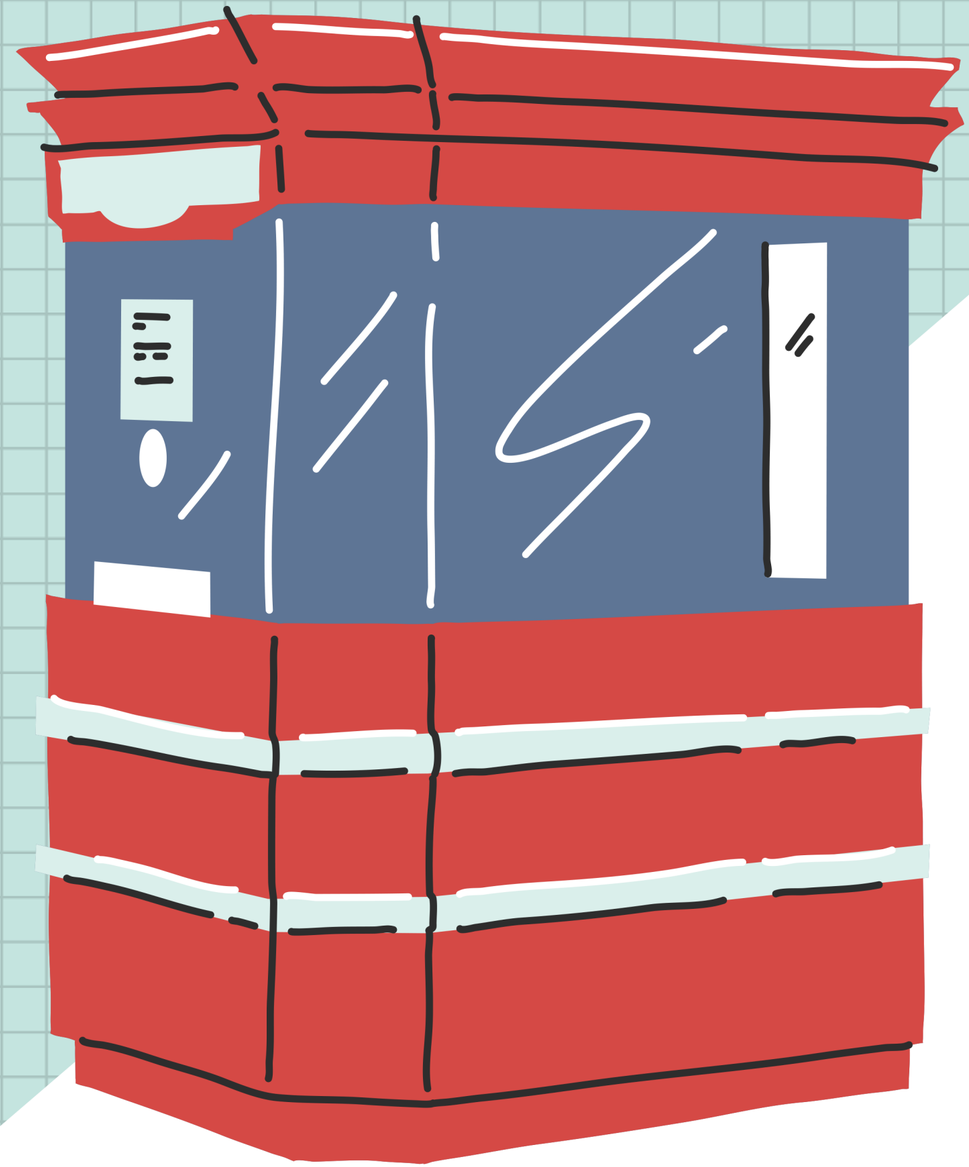
Lord of the Strings

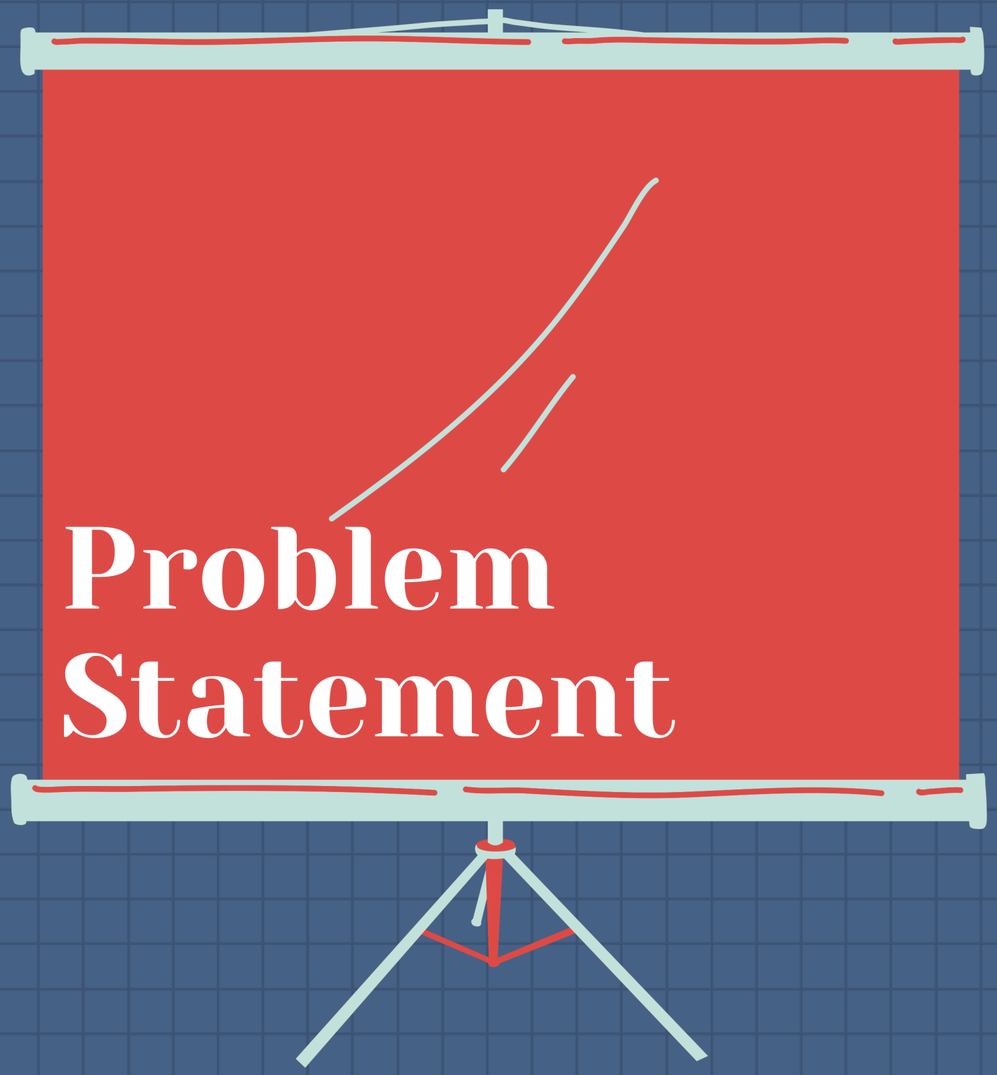
Sadhanha Anand, Rachita Harit, Meghana Kanthadai,

Kumar Kishalaya, Yajur Sehra

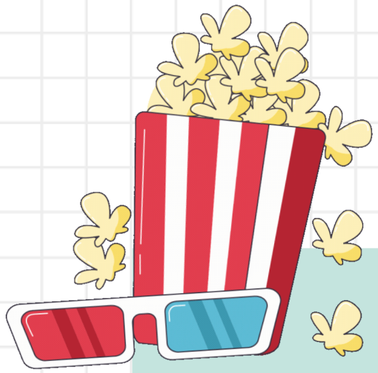
Agenda

1. Problem Statement
2. What are we solving?
3. Solution Approach
4. Predictive Modeling
5. Model Performance
6. Business Impact and Insights
7. Summary





- Movie studios, OTT platforms and Distributors struggle to gauge audience sentiment towards movies
- Traditional methods of sentiment analysis are time-consuming and lack real-time insights.



What are we solving?

Objective

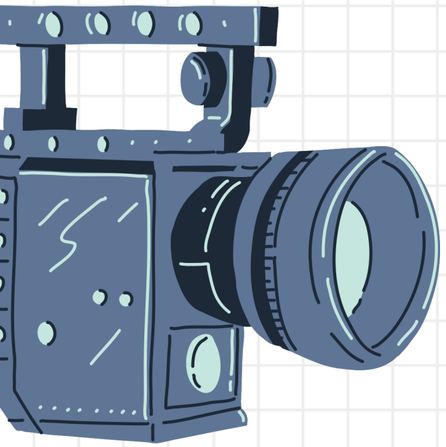
Develop a machine learning model to **Analyze Comments** on movie trailers and predict sentiment as positive or negative.

Key Outcomes

Reduce time and effort with an **Automated Solution**

Provide **Real-Time Sentiment Analysis** to understand audience reactions quickly

Use sentiment data to **Tailor Marketing strategy & PR Campaigns** of movies more effectively



Solution Approach

1

Data Source

IMDB dataset for sentiment analysis (Kaggle IMDB Dataset).
Size: 5k Obsv.

2

Data Preprocessing

Clean and preprocess the dataset to ensure data quality.

3

Vectorization

Text representation techniques:

- Bag of Words
- TF-IDF
- N-gram

4

Model Development

- Logistic Regression
- Support Vector Machine (SVM)
- Random Forest

5

Model Evaluation

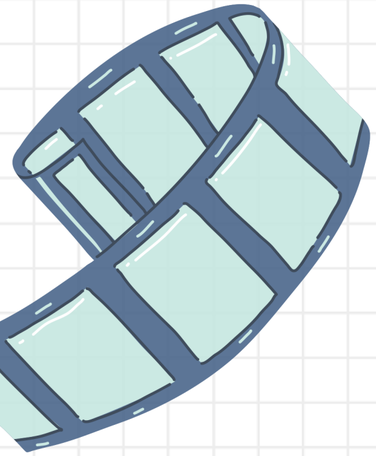
Evaluate the model performance using different metrics

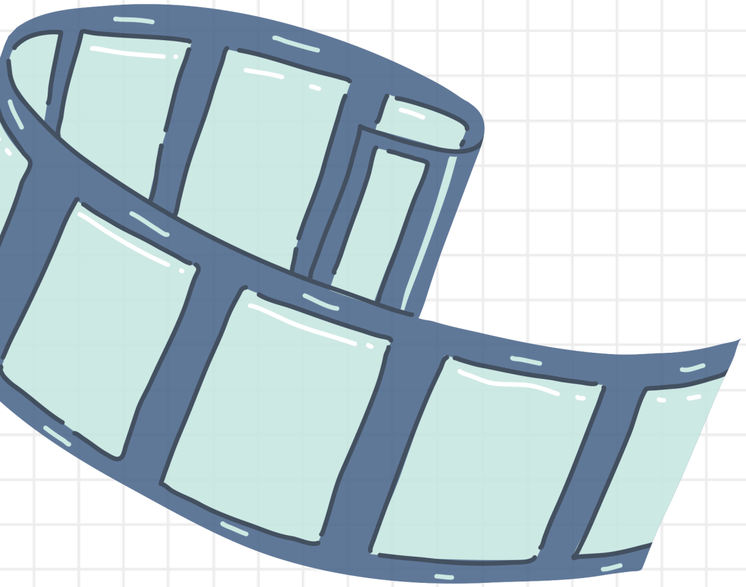
6

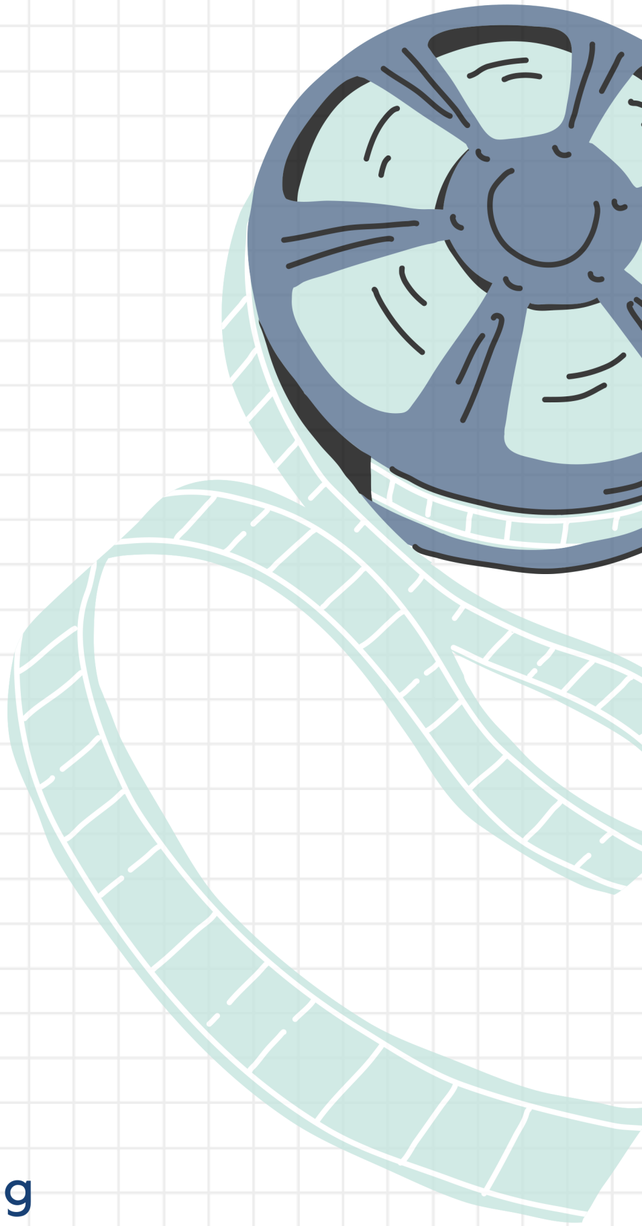
Performance Analysis

Comparing performance of vectorisation methods, with strengths & weaknesses

Embedding Methods







Bag of Words

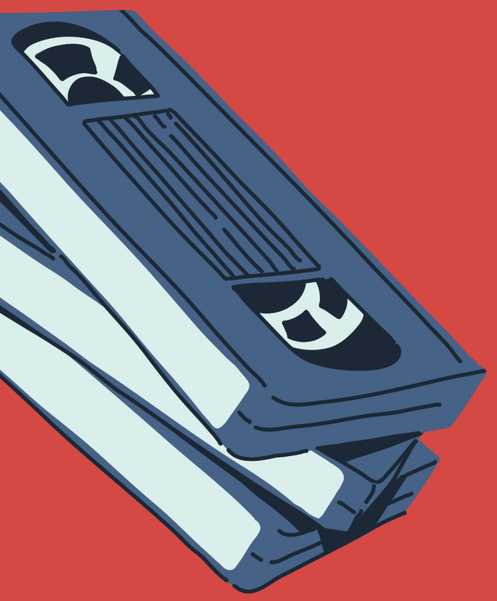
Transform text into fixed-length vectors based on their **frequency**.

TF-IDF

Term Frequency-Inverse Document Frequency to weigh the **importance of words**.

N-gram

Capture context by analyzing **sequences of words** together.



Predictive Modeling

Algorithm

- Logistic Regression
- Support Vector Machine
- Random Forest

Process

- Data preprocessing
- Feature extraction
- Model training
- Evaluation

Tools

- Google Colab
- Python Libraries

Model Evaluation

Bag of Words:

Best Model - Logistic Regression

Accuracy - 0.836

F1 score - 0.84

Precision - 0.84

Recall - 0.84

TF-IDF:

Best Model - Logistic Regression

Accuracy - 0.859

F1 score - 0.86

Precision - 0.86

Recall - 0.86

N-gram:

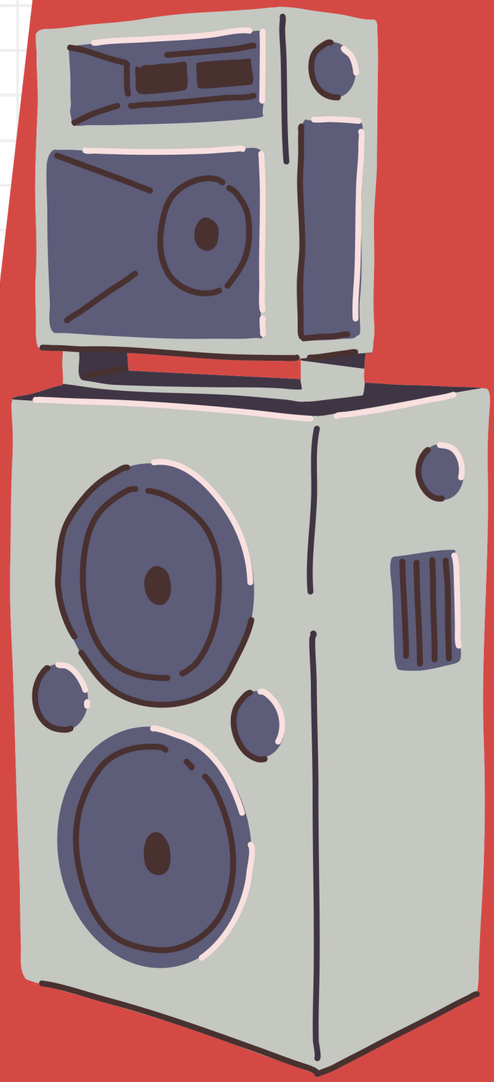
Best Model - Logistic Regression

Accuracy - 0.68

F1 score - 0.69

Precision - 0.7

Recall - 0.69



Model Evaluation

(Logistic Regression)

Bag of Words

	Coefficient
excel	1.244352
favorit	1.207084
today	1.152822
rare	0.979007
hardcor	0.974732
awesom	0.956237
chanc	0.956209
hilari	0.931983
great	0.893248
everyon	0.872530
	Coefficient
worst	-1.859294
wast	-1.795334
bore	-1.321739
wors	-1.239298
terribl	-1.220637
aw	-1.214977
poor	-1.184661
disappoint	-1.140629
fail	-1.070963
lack	-0.998401

TF-IDF

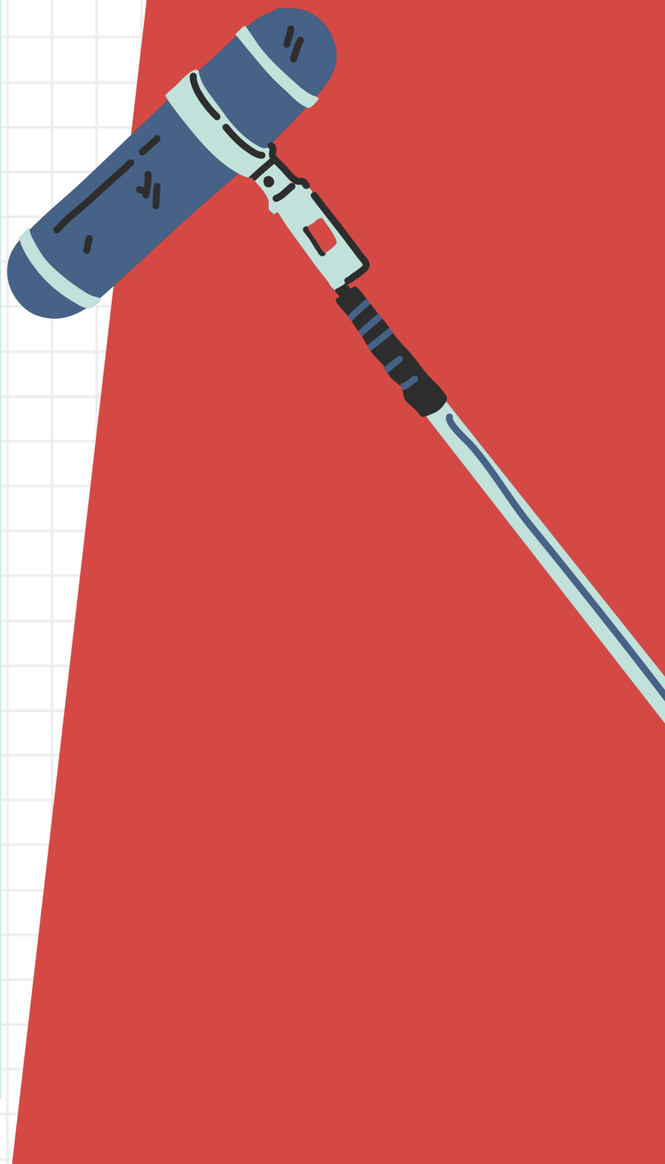
	Coefficient
great	4.526425
love	3.151120
excel	2.857981
best	2.660722
favorit	2.501688
enjoy	2.425616
today	2.151377
beauti	2.103068
well	2.073834
still	1.793836
	Coefficient
bad	-4.828833
worst	-3.868840
wast	-3.611673
bore	-2.845393
aw	-2.556824
poor	-2.505383
terribl	-2.485608
noth	-2.468734
wors	-2.370882
horribl	-2.274261

N-gram

	Coefficient
br highli recommend	1.340411
best movi ever	1.061678
thoroughli enjoy movi	1.021018
highli recommend movi	1.011033
highli recommend film	0.997118
first saw film	0.969146
well worth watch	0.934754
would definit recommend	0.929709
second world war	0.928946
one best movi	0.915763
	Coefficient
worst movi ever	-1.974697
dont wast time	-1.584424
one worst film	-1.478354
worst film ever	-1.371469
mysteri scienc theater	-1.124830
realli want like	-1.120302
realli bad movi	-1.091495
wast time money	-1.017944
complet wast time	-0.992134
br want see	-0.970608

Performance Analysis

Model & Vectorization	Best Use Cases (Examples from ChatGPT)	Key Insights
TF-IDF - Logistic Regression	Ideal for reviews where specific words carry significant weight. Example: "The plot was incredibly intricate and captivating."	High accuracy and AUC. Best for understanding the weight of specific words.
N-Gram - Logistic Regression	Better for texts with sarcasm, idioms, or subtle cues. Example: "Oh, great, another rainy day."	Captures context better than simple keywords, crucial for nuanced sentiment.
Random Forest with Bag of Words or TF-IDF	Great for messy language, like in movie reviews. Example: "The cinematography was stunning, but the plot lacked depth."	Robust for complex, variable reviews. Balances accuracy with feature handling.



Business Impact and Insights

Enhanced Audience Engagement

- Gain insights into audience preferences to refine marketing strategies.
- Implement targeted marketing to enhance positive responses.
- Address criticisms promptly to mitigate negative feedback.

Predictive Insights

- Early indicator of the potential success of a movie, helping studios predict box office performance based on audience reactions to trailers.
- Decide on future investments, like what movies to produce or which marketing channels to focus on

Strategic Decision Making

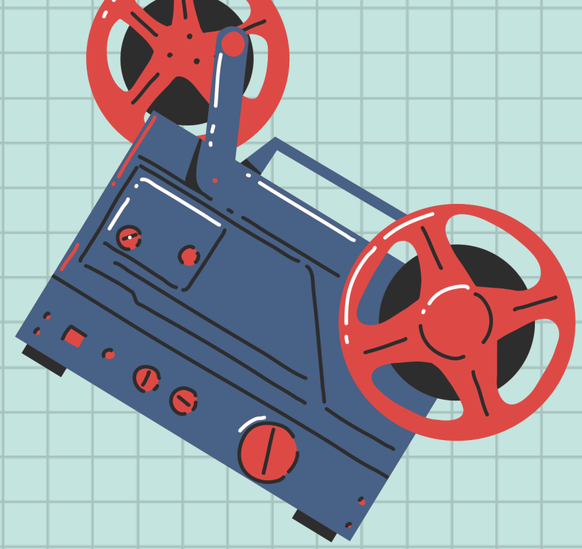
- Assisting OTT platforms in allocating budgets effectively to acquire the rights to the most suitable films.
- Guiding multiple distributors in selecting which shows to run in single screens and multiplexes.
- New projects aligning with audience preferences

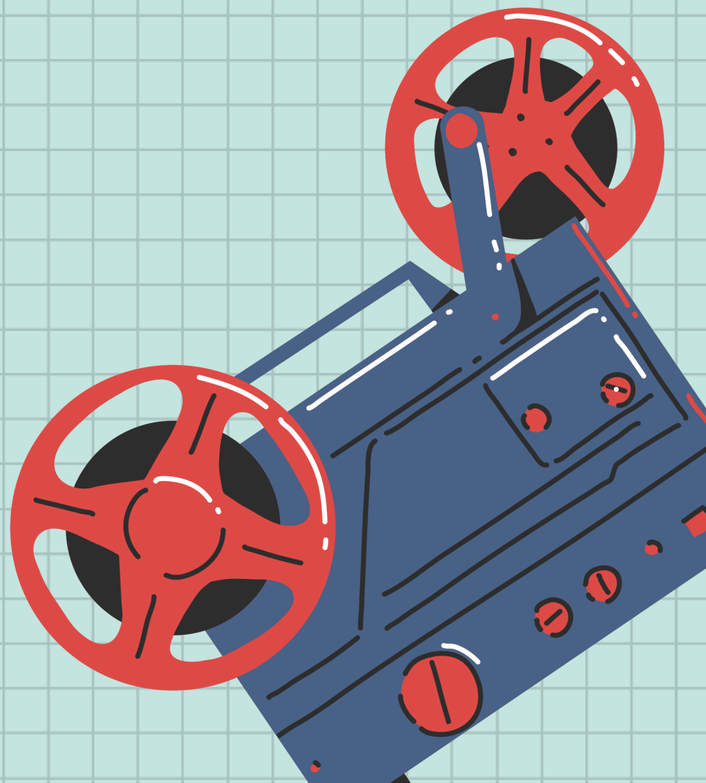
Summary

Developed a ML model to predict sentiment of movie trailer comments and applied data cleaning, text vectorization and binary classification algorithms. Followed by model evaluation to determine effectiveness.

Model Evaluation: **TF-IDF (Logistic Regression)** appears to be the best performer (accuracy of **86%**) out of all the techniques, making it suitable for movie sentiment analysis tasks.

Business Impact:
Audience Engagement , Predictive Insights, Strategic Decisions
Next Steps: Implementation -> Enhancements -> Integration





Thank you!

Questions ?