📊 Predictive Model Summary

🎯 Objective

To predict whether a movie will be a success (positive review) or not (negative review) based on the sentiment of its IMDB user reviews using machine learning models.

📥 Data Used

- Source: IMDB Dataset
- Size: 50,000 labeled reviews (positive or negative)
- Features:
  - review: Text review of a movie
  - sentiment: Label for classification (target variable)

🧹 Preprocessing

- Cleaned HTML tags and special characters
- Lowercased all text
- Tokenized and optionally removed stopwords (depending on method)
- Used TF-IDF Vectorization to convert text into numerical format

🤖 Models Tried

Model	Accuracy	Notes
Logistic Regression	~88%	Lightweight, fast, and interpretable
Random Forest	~85%	Slightly lower performance, slower
Naive Bayes	~84%	Simple and effective for text classification
SVM (Linear)	~89%	Best performer but slower on large text

- Best Model: SVM with TF-IDF features (Accuracy ~89%)

📈 Evaluation Metrics

- Accuracy: Correct predictions out of all predictions
- Precision & Recall: Useful for imbalanced datasets
- Confusion Matrix: Checked for false positives/negatives
- ROC-AUC: Measured ability to distinguish between classes

💡 Insights

- Sentiment polarity is a strong predictor of review outcome.
- VADER compound score correlates well with actual sentiment labels.
- SVM + TF-IDF works best due to its ability to handle high-dimensional sparse data like text.

📦 Deliverables

- Jupyter Notebook with:
  - Data preprocessing
  - Sentiment analysis (VADER)
  - Model training & evaluation
- Visualizations:
  - Confusion matrix
  - Accuracy/ROC curves
  - Genre-wise sentiment bar chart
- README with setup and instructions