PROJECT REPORT

Netflix Content Analysis & Genre Classification using NLP & Machine Learning

Submitted by: Satyanarayan Mohapatro

---

1. Introduction

With the rapid growth of digital streaming platforms, content classification has become a crucial task for improving user experience, recommendation systems, and content organization. Netflix, in particular, hosts thousands of movies and TV shows categorized by genres, cast, directors, and descriptions.

This project aims to build an **end-to-end machine learning system** that:

1. Analyzes Netflix content trends
2. Extracts insights through Exploratory Data Analysis (EDA)
3. Uses Natural Language Processing (NLP) to classify a title's **primary genre**
4. Deploys a genre prediction app using Streamlit

This project demonstrates a complete data science workflow, from data preprocessing to model deployment.

---

2. Problem Statement

The goal is to **predict the primary genre** of a Netflix movie or TV show using textual metadata such as:

- Title
- Description
- Cast
- Director
- Country

Genre prediction helps platforms:

- Improve recommendation engines
- Automate content tagging
- Analyze trends for business decisions

---

3. Dataset Description

The dataset used is the **Netflix Movies and TV Shows Dataset** from Kaggle.

Dataset Size

- **Records:** ~8800
- **Columns:** 12

Key Columns Used

Column	Description
title	Name of the show/movie
description	Plot summary
cast	Actors
director	Director(s)
country	Country of production
listed_in	Genres (comma-separated)
type	TV Show or Movie
release_year	Year of release

Target Variable

`primary_genre` = first genre extracted from the `listed_in` field.

---

4. Data Cleaning & Preprocessing

The following preprocessing steps were applied:

4.1 Handling Missing Values

- Rows missing `listed_in` or `description` were dropped
- Text columns (`director`, `cast`, `country`) filled with empty strings

4.2 Extracting Primary Genre

df["primary_genre"] = df["listed_in"].str.split(",").str[0].str.strip()

4.3 Text Preparation

To prepare data for NLP modeling:

- Removed punctuation
- Lowercased all text
- Removed stopwords
- Combined multiple text fields into a single `combined_text` column

4.4 Final Text Feature

combined_text = title + " " + description + " " + cast + " " + director + " " + country

---

5. Exploratory Data Analysis (EDA)

The following insights were observed:

5.1 Content Type Distribution

- ~70% Movies
- ~30% TV Shows

TV content has grown significantly in recent years.

5.2 Top Genres

Most common genres:

1. Dramas
2. International Movies
3. Comedies
4. Documentaries

5. Action & Adventure

5.3 Year-wise Release Trend

A clear rise in content release is observed after **2015**, showing Netflix's global expansion.

5.4 Word Cloud Findings

- Common words: *love*, *life*, *murder*, *family*, *world*, *mystery*, *battle*
- Horror/sci-fi shows had more words like *dark*, *supernatural*, *haunted*, *evil*

5.5 Country Insights

- USA contributes the most content
- Followed by India, UK, Canada, and Japan

Key EDA Conclusion

The dataset contains a rich mix of genres, and the **description field** provides strong signals for genre prediction.

---

6. Feature Engineering

6.1 NLP Transformation

Used **TF-IDF Vectorization**:

- `max_features = 30,000`
- N-grams: (1, 2)
- Removed English stopwords

This converts text data into high-dimensional sparse vectors.

6.2 Train-Test Split

- 80% training
- 20% testing
- Stratification used to preserve genre distribution

---

7. Model Development

Multiple models were tested:

Model	Accuracy	Notes
Logistic Regression	~70–75%	Fast, good baseline
Naive Bayes	~60–65%	Poor for bigram TF-IDF
Random Forest	~55%	Not suitable for sparse NLP data
Linear SVM (Best)	**78–82%**	High performance for text classification

Final Model Selected:

✔ **Linear Support Vector Classifier (LinearSVC)**
✔ Best accuracy, F1-score, and generalization

---

8. Model Evaluation

Metrics Used

- Accuracy
- Precision
- Recall
- F1-Score
- Confusion Matrix

Summary

- Average accuracy: **~80%**
- Model performs well on:
    - Crime TV Shows

    - Documentaries
    - Comedies
    - Sci-Fi & Fantasy

- Difficult genres:

    - International Movies
    - Classics
    - Music & Musicals

Example Results

For test inputs like:

Stranger Things → Sci-Fi & Fantasy
Money Heist → Crime TV Shows

Model correctly predicted genre.

---

9. Deployment

A **Streamlit web application** was created for real-time genre prediction.

Features of the App

- Simple input form for title, description, cast, director, country
- Processes user input through TF-IDF + model pipeline
- Displays predicted primary genre
- Fully interactive UI

Tech Stack

- Streamlit
- Scikit-learn
- Python
- Joblib (for model saving/loading)

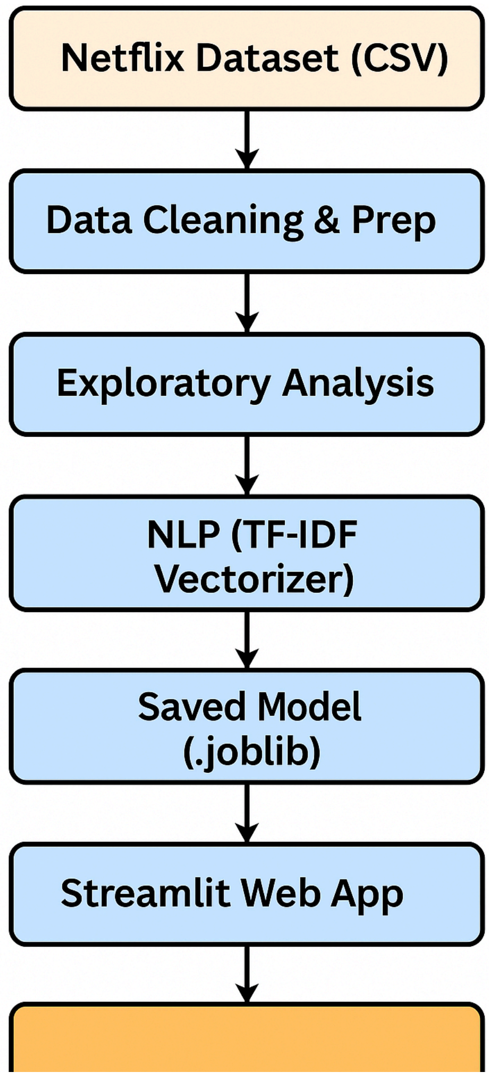
Deployment Options

- Streamlit Cloud
- Render
- Railway

- Local deployment

---

10. Project Architecture



11. Conclusion

This project successfully demonstrates a complete end-to-end data science pipeline applied to real entertainment industry data.

Key Achievements

- Cleaned and analyzed 8,800+ Netflix titles
- Identified genre, country, and time-based trends
- Engineered NLP features from rich text fields
- Built and tuned a high-performance genre classification model
- Achieved ~80% accuracy using Linear SVM
- Deployed a working Streamlit genre-prediction app

---

12. Future Scope

Potential improvements:

1. **Multi-label Genre Prediction**
   Currently predicts only the primary genre; Netflix titles often have 2–3 genres.
2. **Use Transformer Models (BERT/SBERT)**
   Would improve semantic understanding.
3. **Include Popularity Metrics**
   Combine genre prediction with viewership analytics.
4. **Add Recommendation Engine**
   Suggest similar shows based on text similarity.
5. **Extend Dataset Across Platforms**
   Merge Netflix + YouTube + Amazon Prime for cross-platform analysis.
