Movie Genre Classification

Introduction :

This project demonstrates how **Natural Language Processing (NLP)** can be used to predict the **genre of a movie** based solely on its description or synopsis. A machine learning model is trained to classify movies into genres, and the final solution is deployed using **Streamlit** for interactive web-based predictions. This is a practical use case of NLP and multi-class classification in the entertainment industry

---

Dataset Used:

- **Source:** Kaggle
- **Files:**
  - `kaggle_movie_train.csv`
  - `kaggle_movie_test.csv`
- **Key Columns:**
  - `text` (Movie overview or description)
  - `label` (Genre)

Libraries Used :

- **pandas** – Used for loading, cleaning, and managing tabular data.
- **numpy** – Provides support for numerical operations and array handling.
- **nltk** – Helps with natural language processing tasks like removing stopwords.
- **matplotlib** – Used for creating visual plots and charts.
- **seaborn** – Built on top of matplotlib for more attractive statistical visualizations.
- **scikit-learn** – Used for building the machine learning model and evaluating performance.
- **wordcloud** – Generates a visual cloud of the most common words from the text data.
- **streamlit** – Creates a simple and interactive web app for users to input data and see predictions

---

Exploratory Data Analysis (EDA) :

A variety of visualizations were used to understand data trends and distribution:

1. **Genre Distribution:**
   A bar chart shows how many movies belong to each genre.
2. **WordCloud:**
   A WordCloud displays the most common words in all movie descriptions.
3. **Overview Length Distribution:**
   A histogram shows the length of movie descriptions using word count.
4. **Top Words by Genre:**
   Bar charts show the most frequent words for each movie genre.
5. **TF-IDF Word Importance:**
   Top words were selected based on their importance using TF-IDF scores.

6. **Overview Length by Genre:**
   A boxplot compares description lengths across different genres.

Features Used:

- **Text Preprocessing:**
  - Removal of HTML tags and non-alphabet characters
  - Lowercasing
  - Stopword removal using `nltk`

- **Text Vectorization:**
  - TF-IDF with max 5000 features

- **Target Variable:**
  - `genre` (multi-class classification)

Model Training:

- **Split:** 80% training / 20% testing
- **Algorithm Used:** Logistic Regression
- **Pipeline:**

```
Pipeline([
    ('tfidf', TfidfVectorizer(max_features=5000)),
    ('clf', LogisticRegression(max_iter=300))
])
```

- **Evaluation:**
  - Accuracy score
  - Classification report (precision, recall, F1-score)

Model Performance:

- **Accuracy:** (e.g., 85-90%) — exact number printed in notebook
- **Precision/Recall/F1:** Evaluated for each genre class

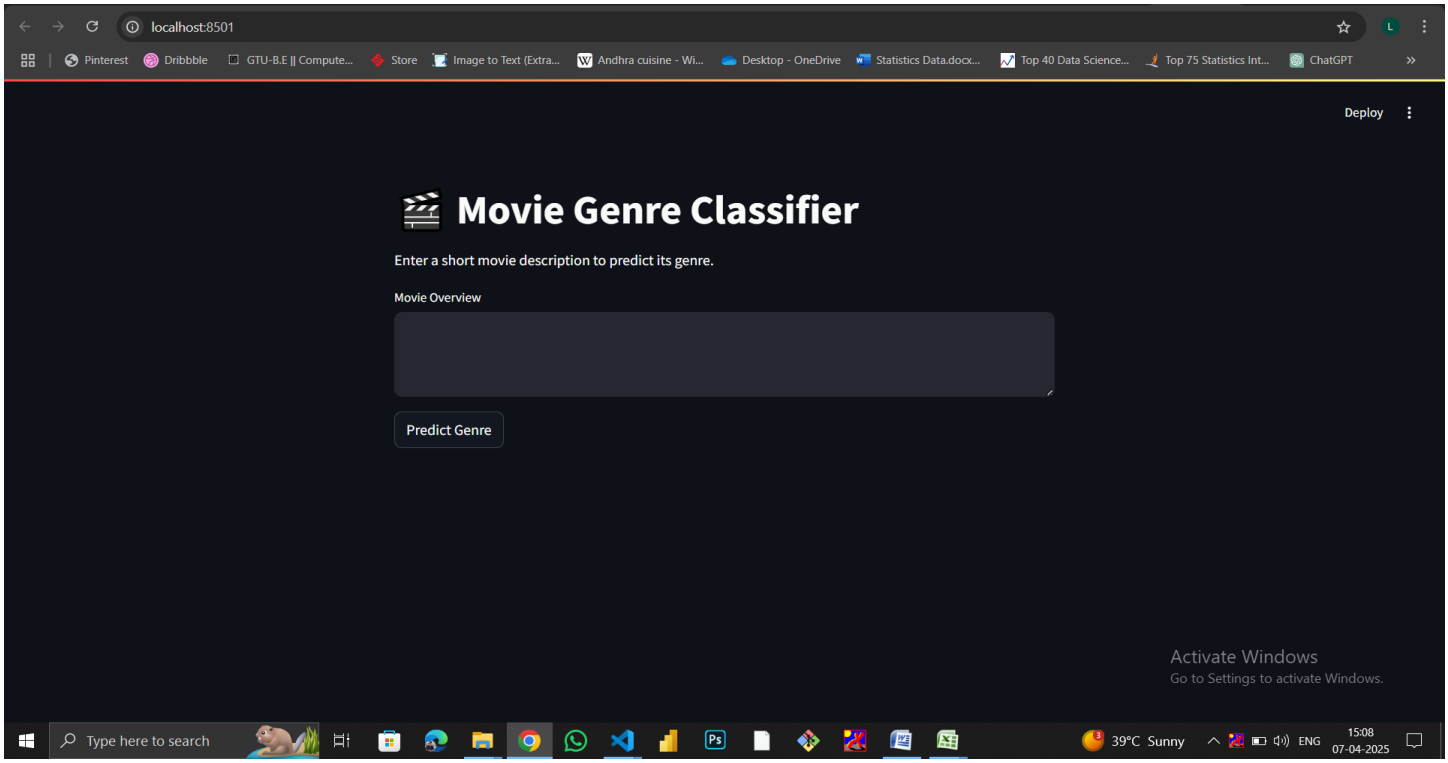
Streamlit Application :

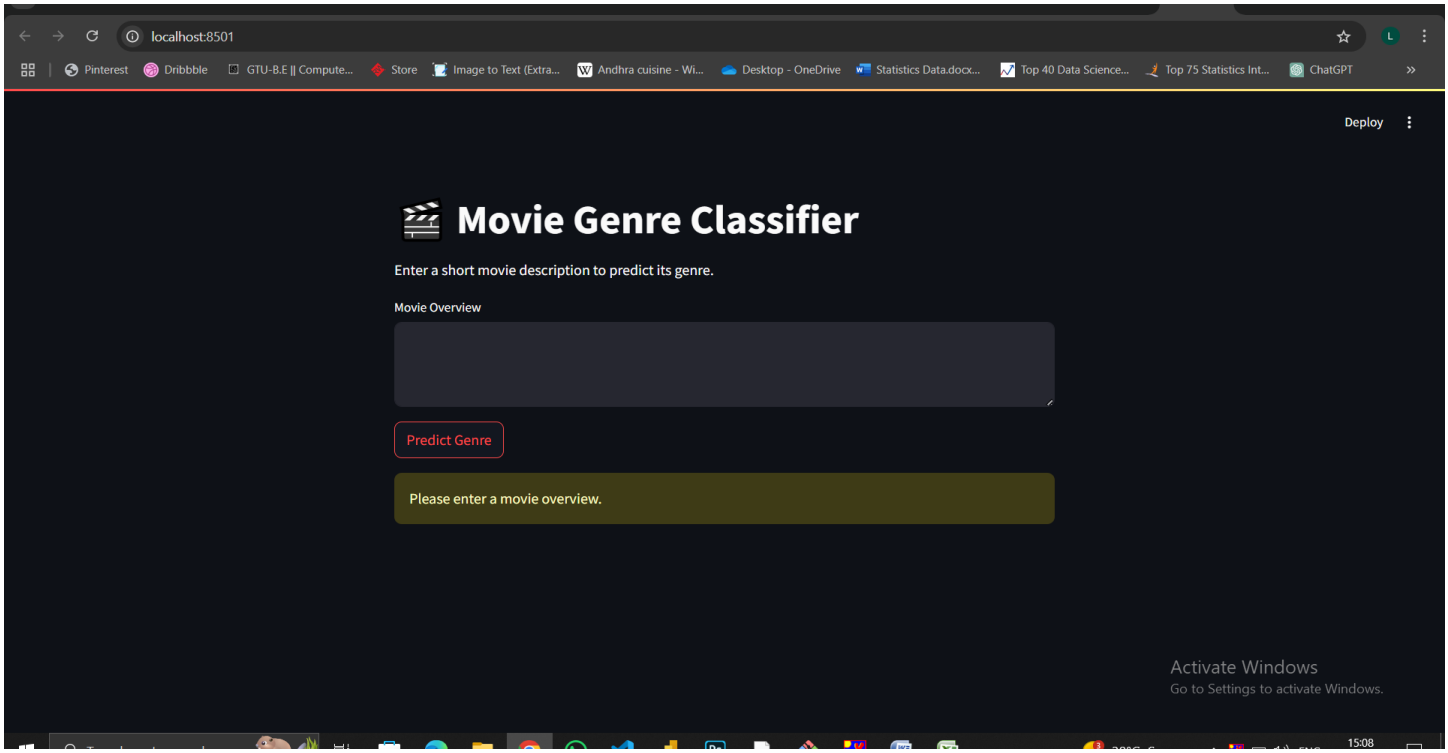
- **Input:** User enters movie description
- **Output:** Predicted genre
- **Model Used:** Saved `genre_classifier.pkl`

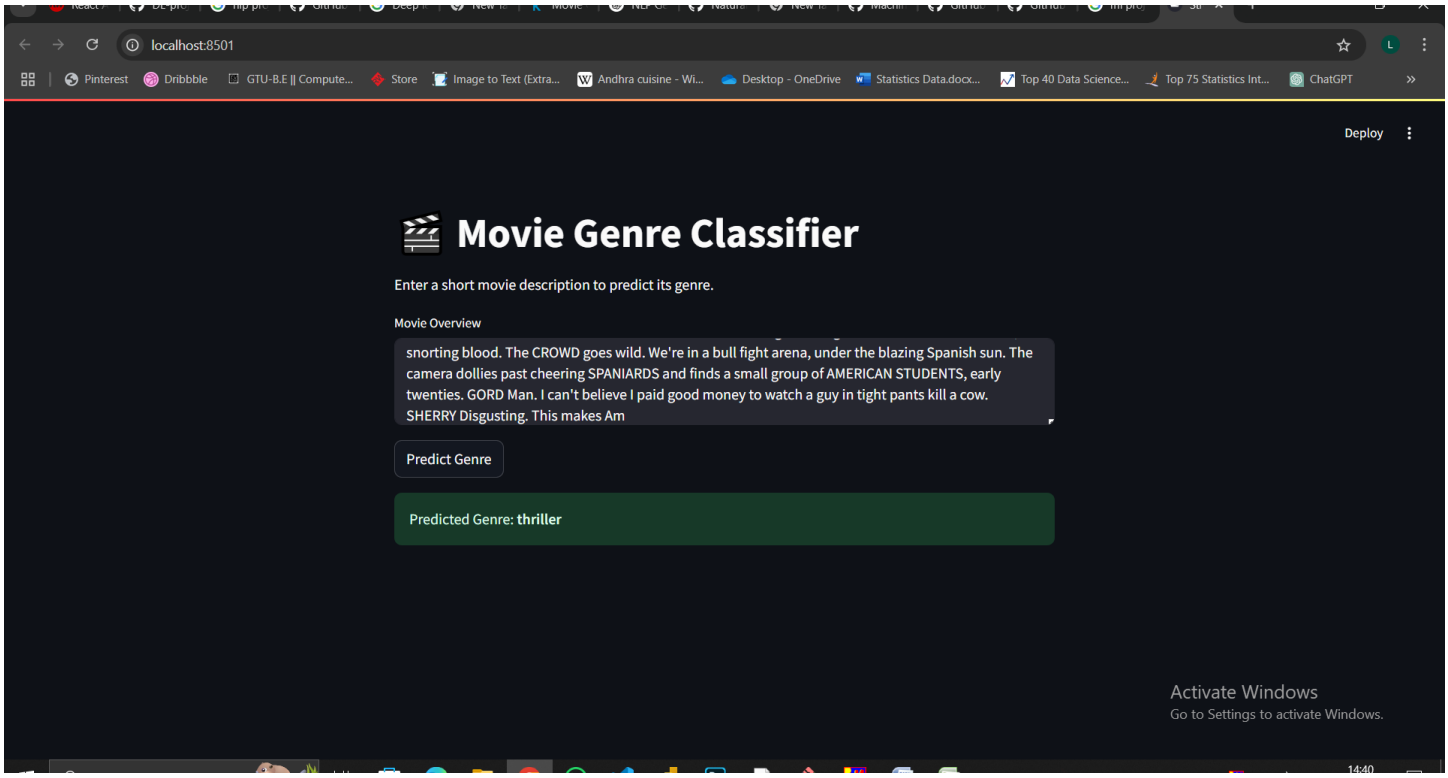
Future Improvements:

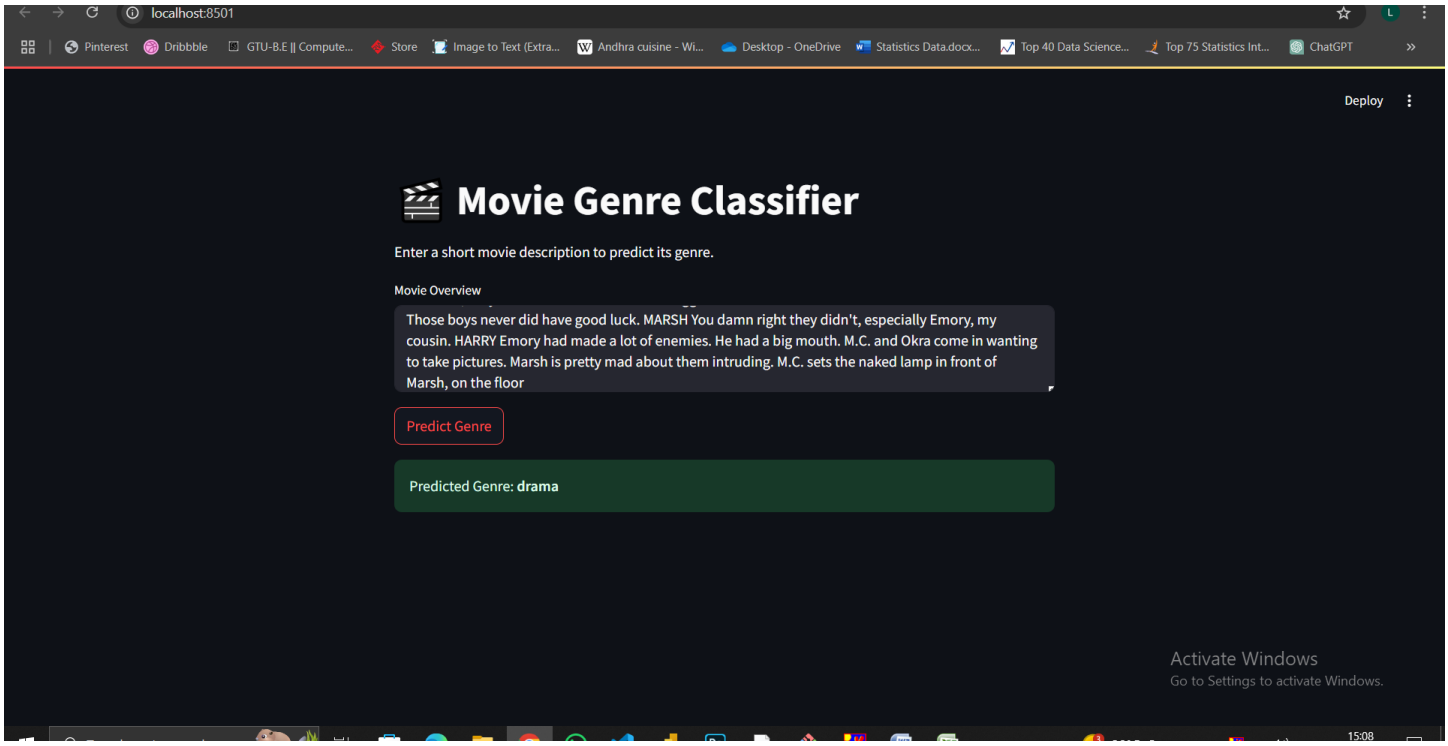
- Use BERT or Transformer models for better accuracy
- Enable multi-label classification

Output Image :






localhost:8501
Deploy
Movie Genre Classifier
Enter a short movie description to predict its genre.
Movie Overview
snorting blood. The CROWD goes wild. We're in a bull fight arena, under the blazing Spanish sun. The camera dollies past cheering SPANIARDS and finds a small group of AMERICAN STUDENTS, early twenties. GORD Man. I can't believe I paid good money to watch a guy in tight pants kill a cow. SHERRY Disgusting. This makes Am
Predict Genre
Predicted Genre: thriller
Activate Windows
Go to Settings to activate Windows.


localhost:8501
Deploy
Movie Genre Classifier
Enter a short movie description to predict its genre.
Movie Overview
Those boys never did have good luck. MARSH You damn right they didn't, especially Emory, my cousin. HARRY Emory had made a lot of enemies. He had a big mouth. M.C. and Okra come in wanting to take pictures. Marsh is pretty mad about them intruding. M.C. sets the naked lamp in front of Marsh, on the floor
Predict Genre
Predicted Genre: drama
Activate Windows
Go to Settings to activate Windows.