Machine learning project

MOVIE GENRE CLASSIFICATION

Dataset Link :

https://www.kaggle.com/datasets/hijest/genre-classification-dataset-imdb
https://www.kaggle.com/datasets/hijest/genre-classification-dataset-imdb
https://www.kaggle.com/datasets/hijest/genre-classification-dataset-imdb
https://www.kaggle.com/datasets/hijest/genre-classification-dataset-imdb

Link of the google collab file :

https://colab.research.google.com/drive/1im6OjSWedMcKzLkkKK02jTm8spSajOf-#scrollTo=dvr2sTW-1hiz

Github Repository Link :

https://github.com/AbhishekYadav-01/Encryptix/tree/main/MOVIE%20GENRE%20CLASSIFICATION

Aim :

Create a machine learning model that can predict the genre of a movie based on its plot summary or other textual information. You can use technique TF-IDF with classifiers such as Naive Bayes, Logistic Regression, or Support Vector Machines.

Sol :

Data Preprocessing :

- Text Cleaning : Removal of stopwords, punctuation, and special characters.
- Tokenization : Splitting sentences into words or tokens.
- TF-IDF Vectorization : Convert text data into numerical features using TF-IDF.

Dataset Overview :

The dataset consists of :

- Number of Training Samples : 9560
- Number of Test Samples : 9822

Genre Distribution in Training Set

The training set includes a variety of movie genres, with the following distribution:

drama	2365
documentary	2307
comedy	1308
short	897

horror	379
thriller	313
action	236
western	194
reality-tv	159
family	138
music	130
adventure	129
romance	114
sci-fi	114
adult	114
animation	85
sport	78
crime	74
talk-show	71
fantasy	65
mystery	56
musical	52
biography	47
...	
game-show	37
news	31
war	22

Model Training:

Implemented several classifiers:

■

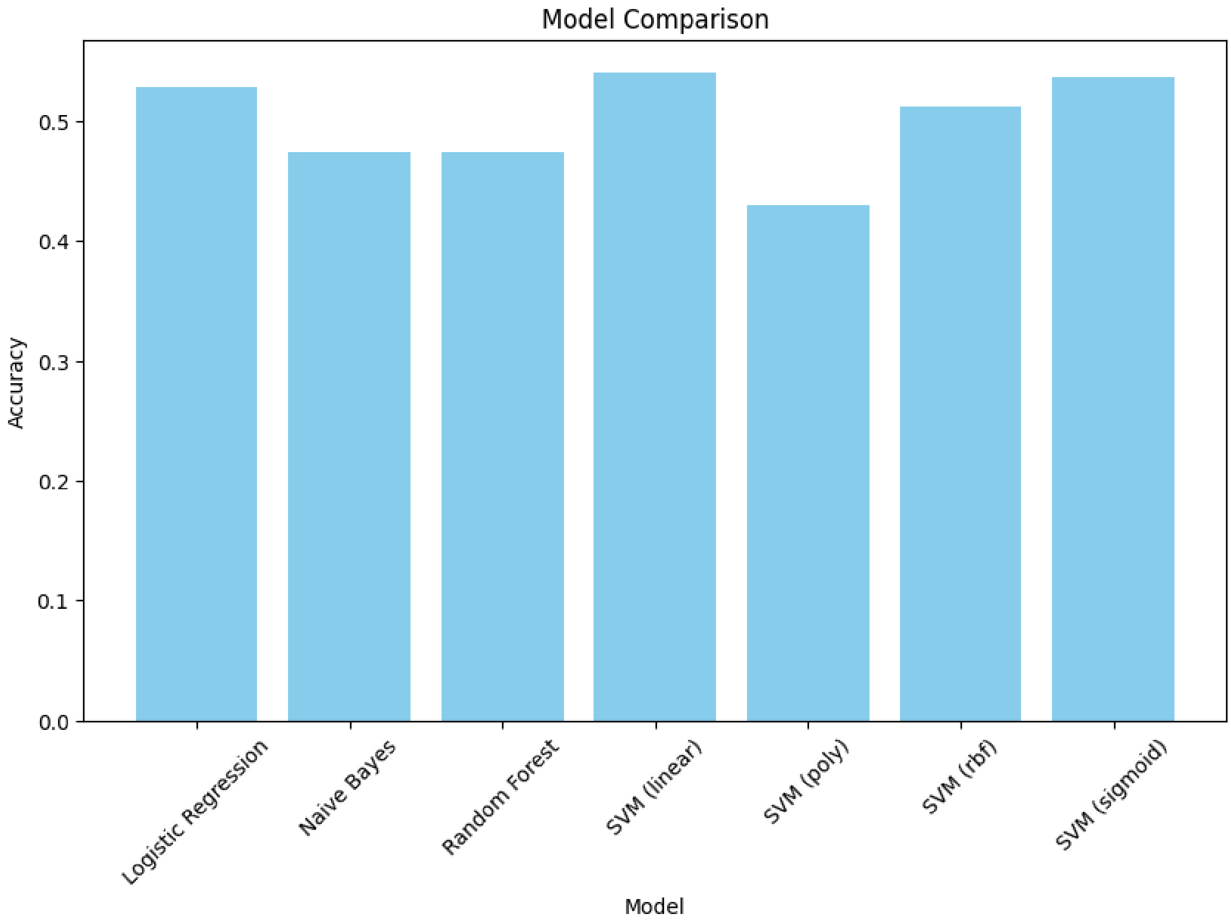
- **Cross-Validation**: Utilized to ensure robustness of models against overfitting.

2. **Model Evaluation:**
   - Metrics used: Accuracy, Precision, Recall, and F1-Score.
   - Confusion matrices were plotted to visualize performance across different genres.

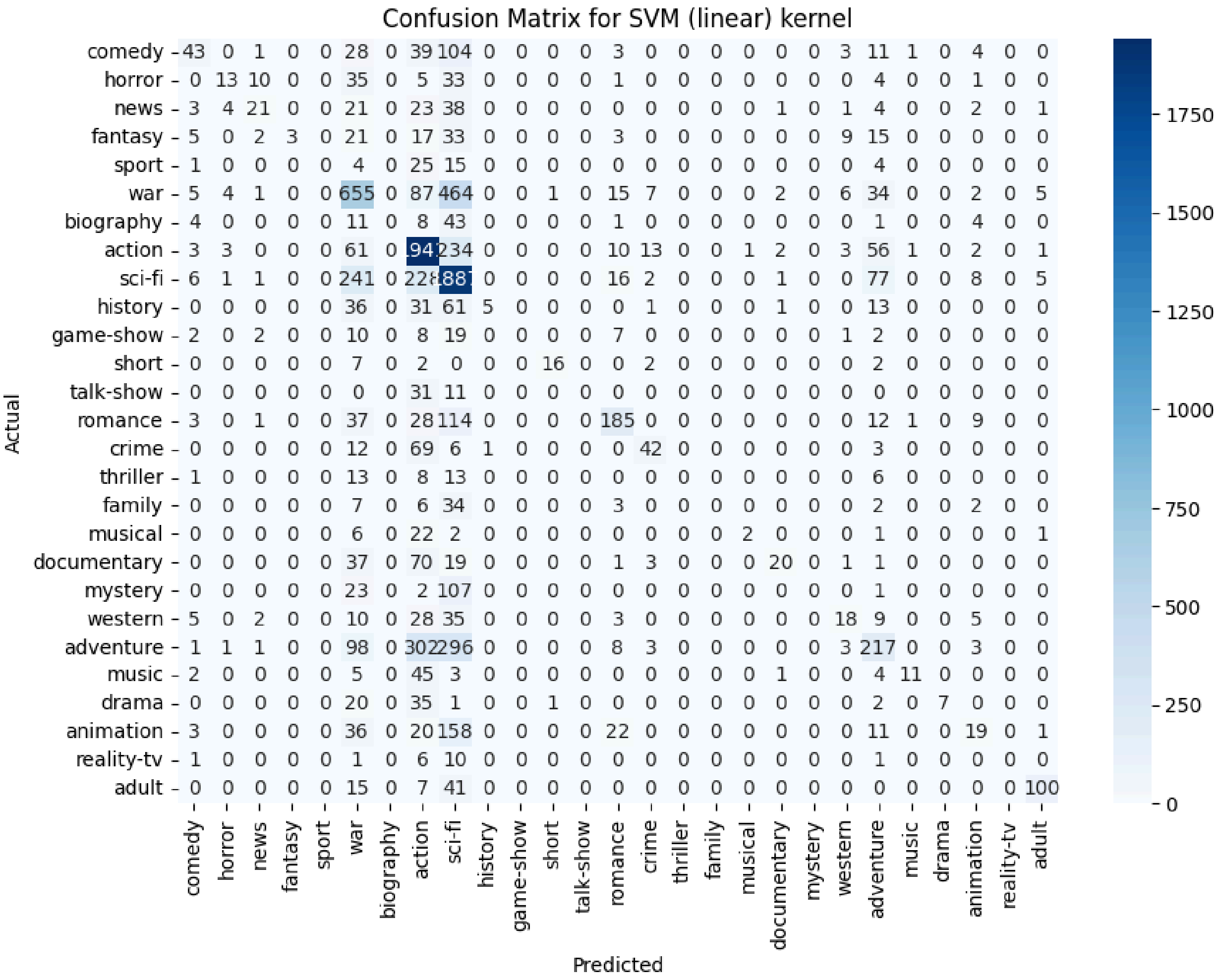
Now we will train all models and see the results on test data:

```
Evaluating Logistic Regression model:
Accuracy (Logistic Regression): 0.5280058193910423
Evaluating Naive Bayes model:
Accuracy (Naive Bayes): 0.47428036994700196

Evaluating Support Vector Machine with linear kernel:
Accuracy (SVM (linear)): 0.5408916138418373


             precision    recall  f1-score   support

      action       0.49      0.18      0.26       237
       adult       0.50      0.13      0.20       102
   adventure       0.50      0.18      0.26       119
   animation       1.00      0.03      0.05       108
   biography       0.00      0.00      0.00        49
      comedy       0.45      0.51      0.48      1288
       crime       0.00      0.00      0.00        72
 documentary       0.63      0.83      0.72      2331
       drama       0.50      0.76      0.60      2473
      family       0.83      0.03      0.06       148
     fantasy       0.00      0.00      0.00        51
   game-show       0.89      0.55      0.68        29
     history       0.00      0.00      0.00        42
      horror       0.67      0.47      0.55       390
       music       0.58      0.32      0.41       133
     musical       0.00      0.00      0.00        41
     mystery       0.00      0.00      0.00        54
        news       0.67      0.06      0.11        34
  reality-tv       0.71      0.13      0.22       152
     romance       0.00      0.00      0.00       133
      sci-fi       0.40      0.16      0.23       115
       short       0.44      0.23      0.30       933
       sport       0.79      0.15      0.26        71
...
weighted avg       0.52      0.54      0.49      9623
```

```
Evaluating Support Vector Machine with poly kernel:
Accuracy (SVM (poly)): 0.42949184246077104
Evaluating Support Vector Machine with rbf kernel:
Accuracy (SVM (rbf)): 0.5128338356022031

Evaluating Support Vector Machine with sigmoid kernel:
Accuracy (SVM (sigmoid)): 0.5370466590460355


Evaluating Random Forest model:
Accuracy (Random Forest): 0.47448820534136965
Model Comparison:
Logistic Regression: 0.5280058193910423
Naive Bayes: 0.47428036994700196
Random Forest: 0.47448820534136965
SVM (linear): 0.5408916138418373
SVM (poly): 0.42949184246077104
SVM (rbf): 0.5128338356022031
SVM (sigmoid): 0.5370466590460355
```



This is the confusion matrix for SVM linear kernal :
(For other models confusion matrix check on github)



Confusion Matrix for SVM (linear) kernel

	comedy	horror	news	fantasy	sport	war	biography	action	sci-fi	history	game-show	short	talk-show	romance	crime	thriller	family	musical	documentary	mystery	western	adventure	music	drama	animation	reality-tv	adult
comedy	43	0	1	0	0	28	0	39	104	0	0	0	0	3	0	0	0	0	0	0	3	11	1	0	4	0	0
horror	0	13	10	0	0	35	0	5	33	0	0	0	0	1	0	0	0	0	0	0	0	4	0	0	1	0	0
news	3	4	21	0	0	21	0	23	38	0	0	0	0	0	0	0	0	0	1	0	1	4	0	0	2	0	1
fantasy	5	0	2	3	0	21	0	17	33	0	0	0	0	3	0	0	0	0	0	0	9	15	0	0	0	0	0
sport	1	0	0	0	0	4	0	25	15	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0
war	5	4	1	0	0	655	0	87	464	0	0	1	0	15	7	0	0	0	2	0	6	34	0	0	2	0	5
biography	4	0	0	0	0	11	0	8	43	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	4	0	0
action	3	3	0	0	0	61	0	194	234	0	0	0	0	10	13	0	0	1	2	0	3	56	1	0	2	0	1
sci-fi	6	1	1	0	0	241	0	228	1887	0	0	0	0	16	2	0	0	0	1	0	0	77	0	0	8	0	5
history	0	0	0	0	0	36	0	31	61	5	0	0	0	0	1	0	0	0	1	0	0	13	0	0	0	0	0
game-show	2	0	2	0	0	10	0	8	19	0	0	0	0	7	0	0	0	0	0	0	1	2	0	0	0	0	0
short	0	0	0	0	0	7	0	2	0	0	0	16	0	0	2	0	0	0	0	0	0	2	0	0	0	0	0
talk-show	0	0	0	0	0	0	0	31	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
romance	3	0	1	0	0	37	0	28	114	0	0	0	0	185	0	0	0	0	0	0	0	12	1	0	9	0	0
crime	0	0	0	0	0	12	0	69	6	1	0	0	0	0	42	0	0	0	0	0	0	3	0	0	0	0	0
thriller	1	0	0	0	0	13	0	8	13	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0
family	0	0	0	0	0	7	0	6	34	0	0	0	0	3	0	0	0	0	0	0	0	2	0	0	2	0	0
musical	0	0	0	0	0	6	0	22	2	0	0	0	0	0	0	0	0	2	0	0	0	1	0	0	0	0	1
documentary	0	0	0	0	0	37	0	70	19	0	0	0	0	1	3	0	0	0	20	0	1	1	0	0	0	0	0
mystery	0	0	0	0	0	23	0	2	107	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
western	5	0	2	0	0	10	0	28	35	0	0	0	0	3	0	0	0	0	0	0	18	9	0	0	5	0	0
adventure	1	1	1	0	0	98	0	302	296	0	0	0	0	8	3	0	0	0	0	0	3	217	0	0	3	0	0
music	2	0	0	0	0	5	0	45	3	0	0	0	0	0	0	0	0	0	1	0	0	4	11	0	0	0	0
drama	0	0	0	0	0	20	0	35	1	0	0	1	0	0	0	0	0	0	0	0	0	2	0	7	0	0	0
animation	3	0	0	0	0	36	0	20	158	0	0	0	0	22	0	0	0	0	0	0	0	11	0	0	19	0	1
reality-tv	1	0	0	0	0	1	0	6	10	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
adult	0	0	0	0	0	15	0	7	41	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100

Conclusion :

In this project, we implemented and evaluated various machine learning models to classify movie genres based on plot summaries. The Support Vector Machine with a linear kernel achieved the highest accuracy (0.541), outperforming other models. Logistic Regression and SVM with a sigmoid kernel also performed relatively well. However, it is important to note that while accuracy is a useful metric, other factors such as precision, recall, and the confusion matrix should also be considered to understand the models' performance better. Future improvements could include exploring deep learning models, optimizing hyperparameters, and incorporating additional features to enhance the classification accuracy.