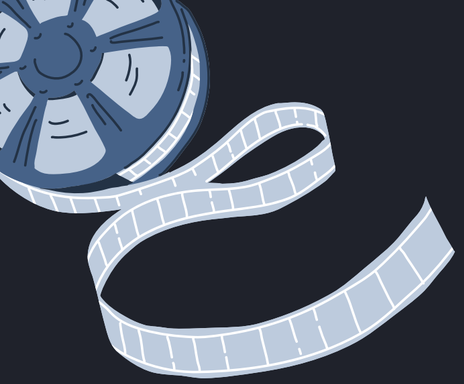
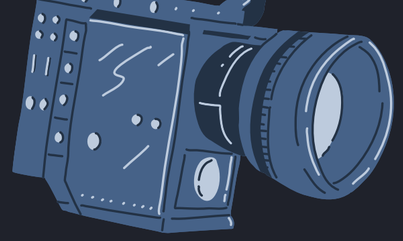
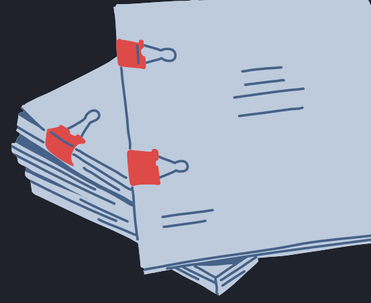


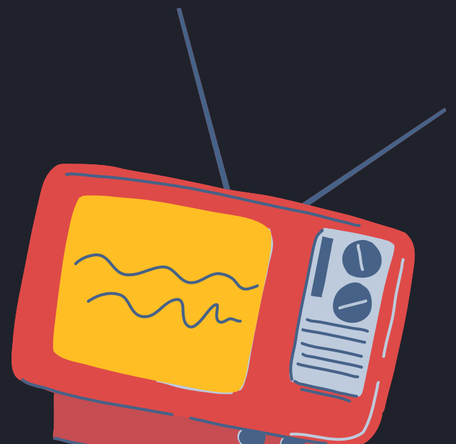



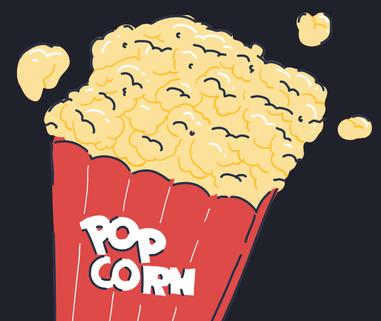


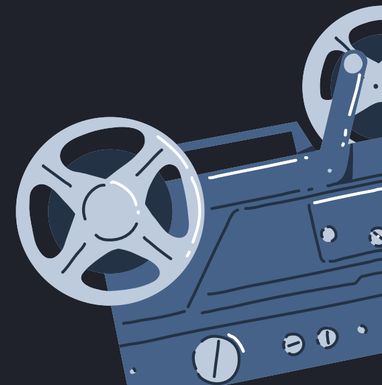
PHÂN LOẠI ĐÁNH GIÁ PHIM

Trí tuệ nhân tạo



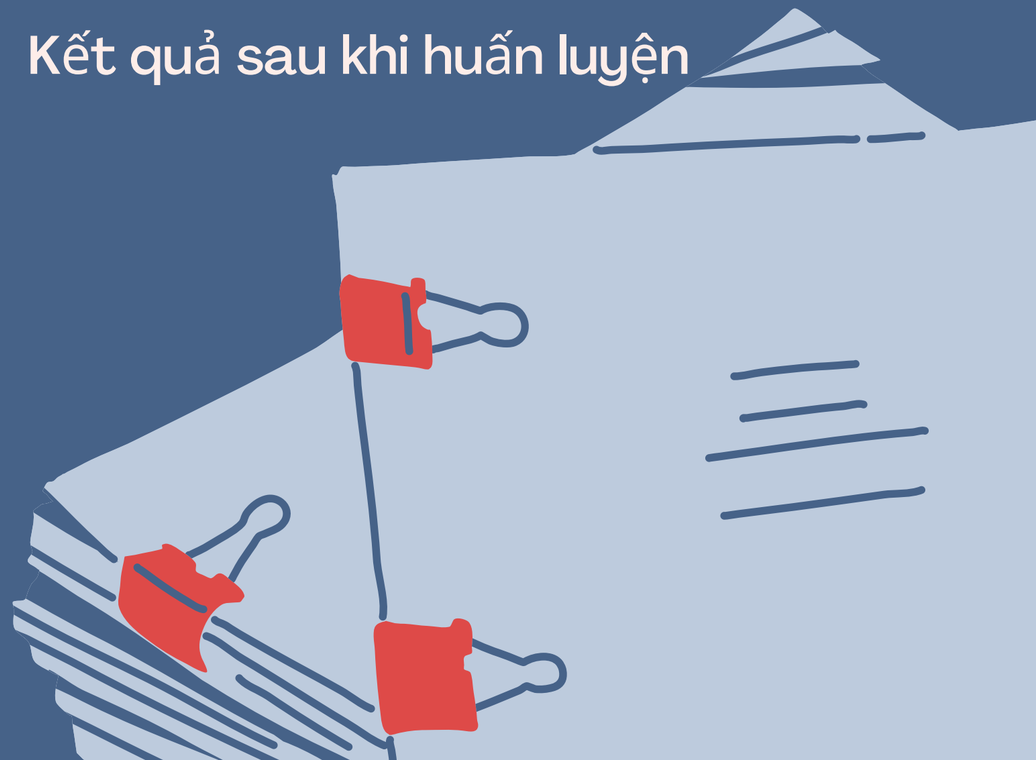







NỘI DUNG

1. Giới thiệu dự án
2. Thông tin bộ dữ liệu
3. Khám phá và trực quan dữ liệu
4. Tiền xử lý dữ liệu
5. Huấn luyện mô hình
6. Kết quả sau khi huấn luyện



1. GIỚI THIỆU DỰ ÁN

MỤC TIÊU ĐỀ TÀI

Đề tài tập trung nghiên cứu và xây dựng mô hình Trí tuệ nhân tạo để phân loại cảm xúc trong các đánh giá phim, nhằm tự động nhận biết nhận xét tích cực hay tiêu cực. Trong bối cảnh lượng đánh giá người dùng tăng nhanh, việc phân tích thủ công trở nên khó khăn, vì vậy các kỹ thuật học máy và xử lý ngôn ngữ tự nhiên (NLP) được áp dụng để xử lý dữ liệu hiệu quả hơn. Đề tài tìm hiểu các bước tiền xử lý văn bản, sử dụng các mô hình như TF-IDF kết hợp thuật toán phân loại, và đánh giá hiệu quả dựa trên các chỉ số Accuracy hay Precision. Kết quả giúp cải thiện khả năng phân tích phản hồi người dùng và hỗ trợ tốt hơn cho hệ thống gợi ý phim.

THÔNG TIN BỘ DỮ LIỆU

NGUỒN DỮ LIỆU

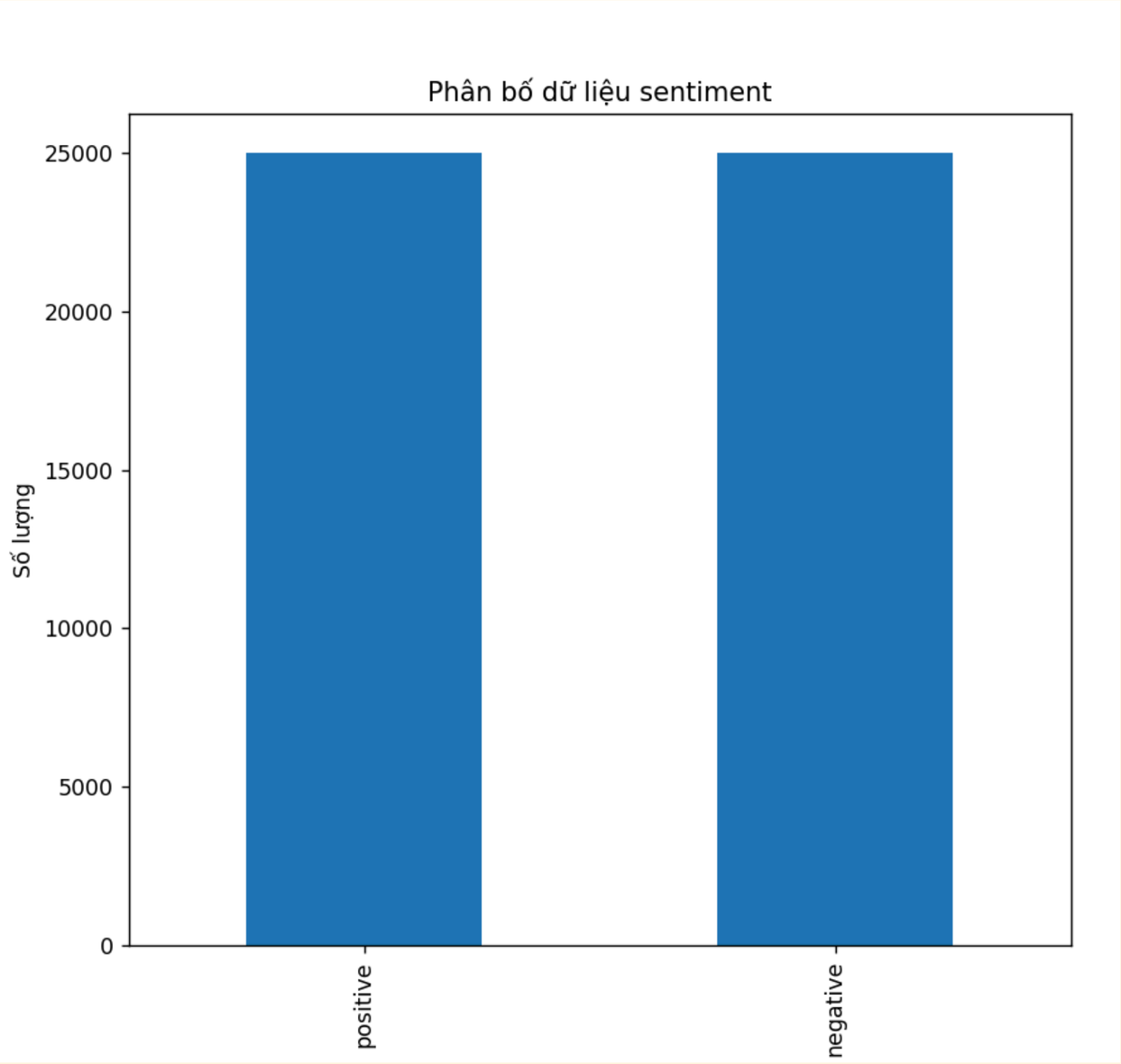
https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews

49582 unique values	2 unique values
One of the other reviewers has mentioned that after watching just 1 Oz episode you'll be hooked. The...	positive
A wonderful little production. The filming technique is very unassuming- very old-time-B...	positive

(50000, 2)

49582 unique values	2 unique values	
One of the other reviewers has mentioned that after watching just 1 Oz episode you'll be hooked. The...	positive	Review : Đánh giá của người xem phim Sentiment : Phân loại đánh giá
A wonderful little production. The filming technique is very unassuming- very old-time-B...	positive	

KHÁM PHÁ VÀ TRỰC QUAN DỮ LIỆU



TOÀN BỘ DỮ LIỆU

Phân bố dữ liệu cân bằng với:

- 25000 review positive (đánh giá tốt)
- 25000 review negative (đánh giá xấu)

TRỰC QUAN DỮ LIỆU

Positive Reviews

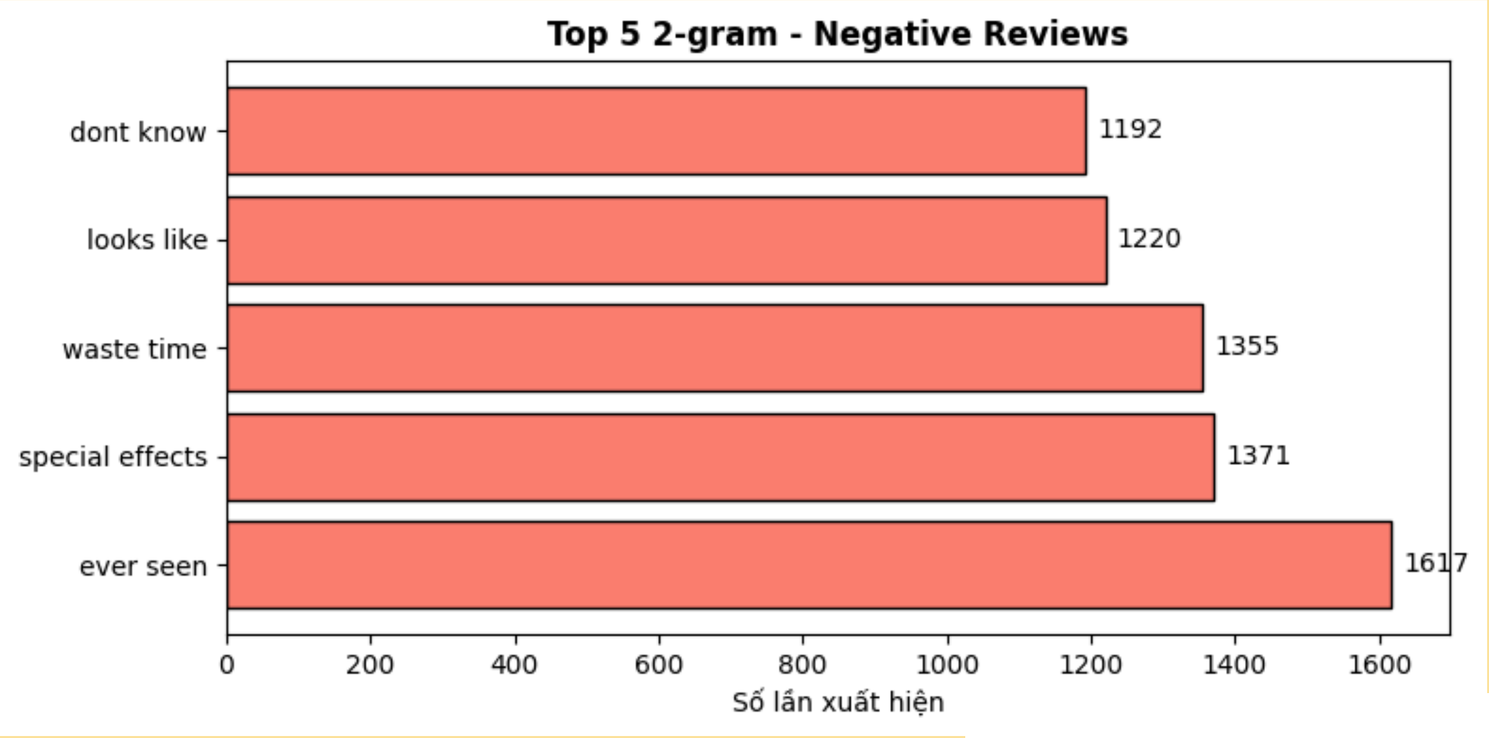
Negative Reviews

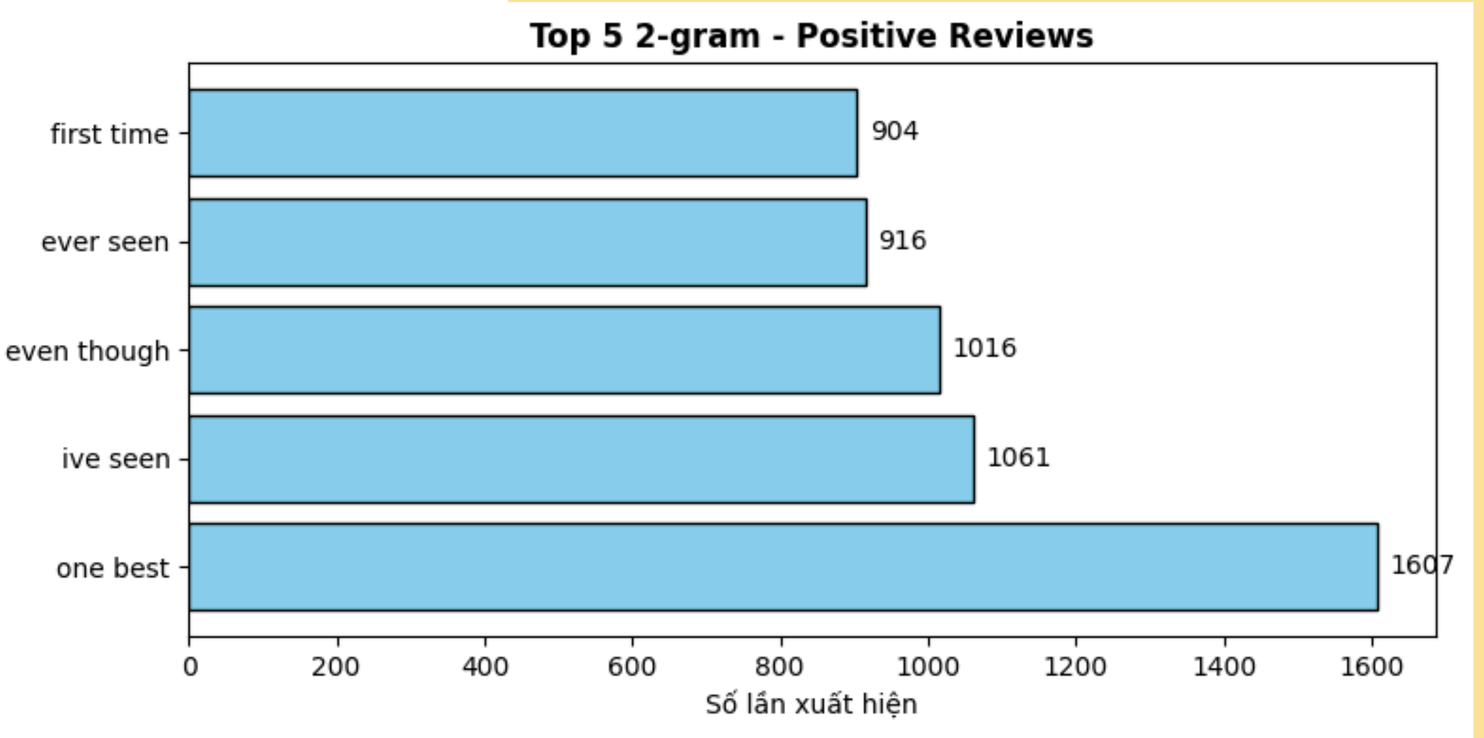
POSITIVE

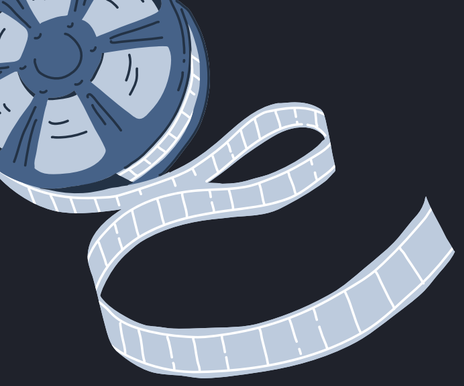
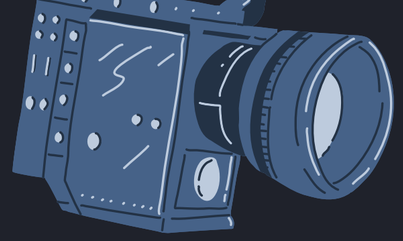
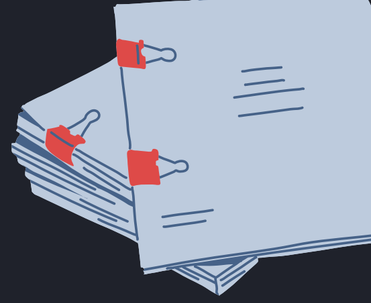
```
Top Positive words:
film
great
good
story
love
best
films
life
think
seen
```

NEGATIVE

```
Top Negative words:
movie
like
bad
really
dont
time
movies
acting
plot
make
```

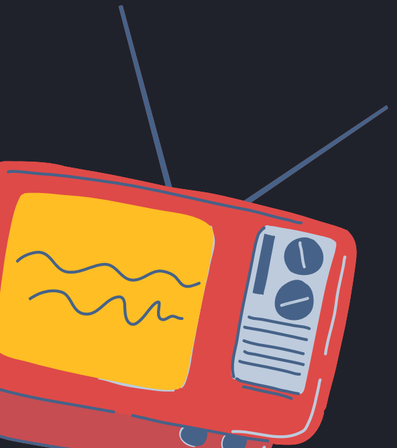

Top 5 2-gram - Negative Reviews
dont know
looks like
waste time
special effects
ever seen
1192
1220
1355
1371
1617
0
200
400
600
800
1000
1200
1400
1600
Số lần xuất hiện

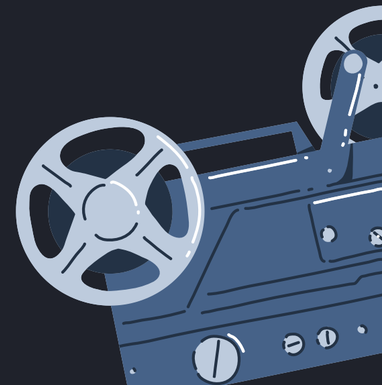

Top 5 2-gram - Positive Reviews
first time
ever seen
even though
ive seen
one best
904
916
1016
1061
1607
0
200
400
600
800
1000
1200
1400
1600
Số lần xuất hiện

TIỀN XỬ LÝ

Dữ Liệu





Kiểm tra missing values

```python
import pandas as pd
import matplotlib.pyplot as plt

df = pd.read_csv("L:/Learning/PYthon/AI/IMDB.csv")

# Kiểm tra tổng số missing values theo từng cột
print(df.isnull().sum())

# Kiểm tra tổng số dòng có dữ liệu thiếu
print("Tổng số dòng missing:", df.isnull().any(
```

```
review       0
sentiment    0
dtype: int64
Tổng số dòng missing: 0
```

Làm sạch dữ liệu

review	sentiment
is called OZ as that is the nickname given to the Oswald	positive
ching but it is a terrificly written and performed piece. A m	positive
he style many of us have grown to love. This v	positive
drama! As a drama the movie is watchable. Parents are c	negative
ach one is connected in one way, or another to the next p	positive
ne of her very few truly sympathetic roles, is a delight. Th	positive
for a world of under water adventure.Oh by the way thank	positive
ildly entertaining respite of the guest-hosts, this show pro	negative
played no less than four times). The film looks cheap and	negative
s movie. If you are young or old then you will love this mo	positive
anymore. Its a low budget film (thats never a p	negative
y resolution when the monster died in the end. I didn't ca	negative
y on the set Mr Boll invites three of his countrymen to pla	negative
owdler (hence bowdlerization) tried to do something simil	negative
a fan but this roll is not bad. Another good thing about th	positive
rilliant insights, just stilted and quite ridiculous (but lots of	negative
g the right feel to the character of Luke. But the major faul	positive
  >The ghost scene at the end was stolen from	negative
ance of Grizzly Adams actor Dan Haggery i think one of th	positive
y bizarre -- by all accounts the true story of this artist woul	negative
acting skills. All in all it's full of excuses to dismiss the film	positive
...almost. The movie is just packed full of crap	negative
on his shoulders and although there isn't anything more of	positive
ill. The animation is VERY bad & cheap, even	negative
because the movie is so ridiculous and predictable. The	negative

cleanreview	sentiment
ristians italians irish moreso scuffles d	positive
written performed piece masterful prod	positive
n still fully control style many us grown	positive
riller drama drama movie watchable pa	negative
phisticated luxurious look taken see p	positive
yes bette davis one truly sympathetic r	positive
pace tv would work world water advent	positive
ning respite guesthosts show probably	negative
ame country tune played less four times	negative
hter like movie young old love movie h	positive
anymoreits low budget film thats never	negative
professor happy resolution monster die	negative
schweiger yes carver german hail brat	negative
ev bowdler hence bowdlerization tried	negative
fan roll bad another good thing movie	positive
ant insights stilted quite ridiculous lots s	negative
haracter luke major fault version straye	positive
ider machine beginning exactly like froc	negative
rance grizzly adams actor dan haggery	positive
e simply bizarre accounts true story ar	negative
nty thrills unintentionally plenty laughsy	positive
packed full crappy oneliners respectab	negative
ries movie shoulders although isnt anyt	positive
ction animation unorganized dialogue v	negative
dge movie ridiculous predictable main	negative

HUẤN LUYỆN MÔ HÌNH & KẾT QUẢ

Phân chia dữ liệu

Tỷ lệ phân chia dữ liệu:
Tập training: 80%
Tập test: 20%

Kích thước tập Train:
X_train: (40000,)
y_train: (40000,)

Kích thước tập Test:
X_test: (10000,)
y_test: (10000,)

Logistic Regression

```
===== LOGISTIC REGRESSION =====
              precision    recall  f1-score   support

    negative       0.91      0.90      0.90      4961
    positive       0.90      0.92      0.91      5039

    accuracy                           0.91     10000
   macro avg       0.91      0.91      0.91     10000
weighted avg       0.91      0.91      0.91     10000

Confusion Matrix:
 [[4451  510]
 [ 427 4612]]
Accuracy: 0.9063
AUC: 0.9062955550061266
```

SVM

```
===== SVM (LinearSVC) =====
              precision    recall  f1-score   support

    negative       0.90      0.90      0.90      4961
    positive       0.90      0.91      0.90      5039

    accuracy                           0.90     10000
   macro avg       0.90      0.90      0.90     10000
weighted avg       0.90      0.90      0.90     10000

Confusion Matrix:
 [[4452  509]
 [ 468 4571]]
Accuracy: 0.9023
AUC: 0.9654622187213869
```

Decision Tree

```
===== DECISION TREE =====
              precision    recall  f1-score   support

    negative       0.74      0.71      0.72      4961
    positive       0.73      0.75      0.74      5039

    accuracy                           0.73     10000
   macro avg       0.73      0.73      0.73     10000
weighted avg       0.73      0.73      0.73     10000

Confusion Matrix:
 [[3524 1437]
 [1244 3795]]
Accuracy: 0.7319
AUC: 0.7317331386441555
```

So Sánh

Mô hình	Accuracy	Precision	Recall	F1-Score
Logistic Regression	**90.26%**	0.90	0.90	0.90
SVM	**90.21%**	0.90	0.90	0.90
Decision Tree	**72.51%**	0.73	0.73	0.72

THANK YOU