UGANDA MARTYRS UNIVERSITY

FACULTY OF SCIENCE

MASTER OF SCIENCE IN INFORMATION SYSTEMS

INTELLIGENT SYSTEMS

MIS 6317

MUGASHA BLAISE BRADLEY

REG NO: 2024-M132-23976

Full Project Documentation: IMDb Sentiment Analysis + RL Model Selector

Overview

This project performs sentiment classification on the IMDb 50,000review dataset using two independent models:

1. **Logistic Regression (TFIDF features)**
2. **LSTM Neural Network (Tokenized & Padded Sequences)**

It then builds a **Reinforcement Learning (RL) Model Selector** that learns when to choose ML or DL for each review to maximize accuracy.

The project is divided into:

- **Part A**: NLP preprocessing and feature engineering
- **Part B**: Two sentiment classifiers
- **Part C**: RL selector using QLearning
- **Part D**: Interpretation and insights

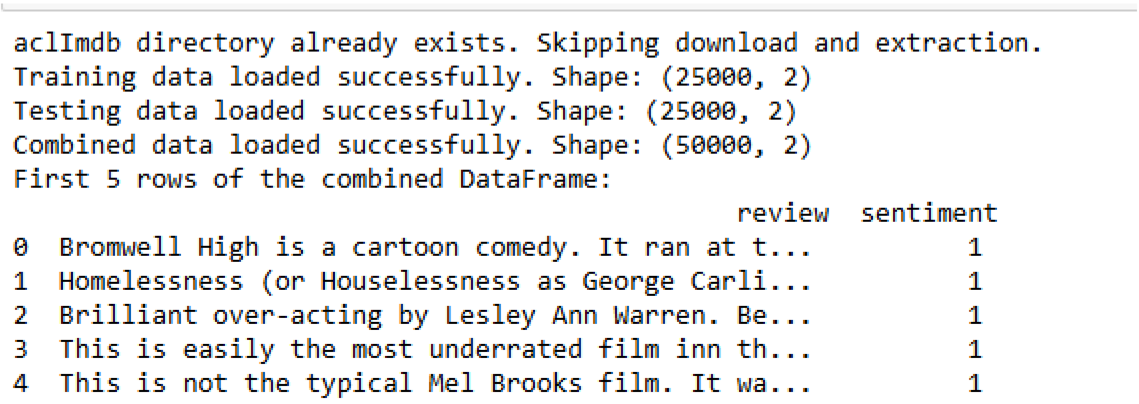
Part A NLP Preprocessing and Feature Engineering

1. Dataset Loading

- 50,000 IMDb reviews loaded from the aclImdb directory.
- Dataset includes:
  - 25,000 positive reviews
  - 25,000 negative reviews
- Combined into a single DataFrame of shape **(50,000, 2)**.

2. Data Inspection

- Dataset is balanced (25k positive, 25k negative).
- Both training and testing reviews merged into one DataFrame.



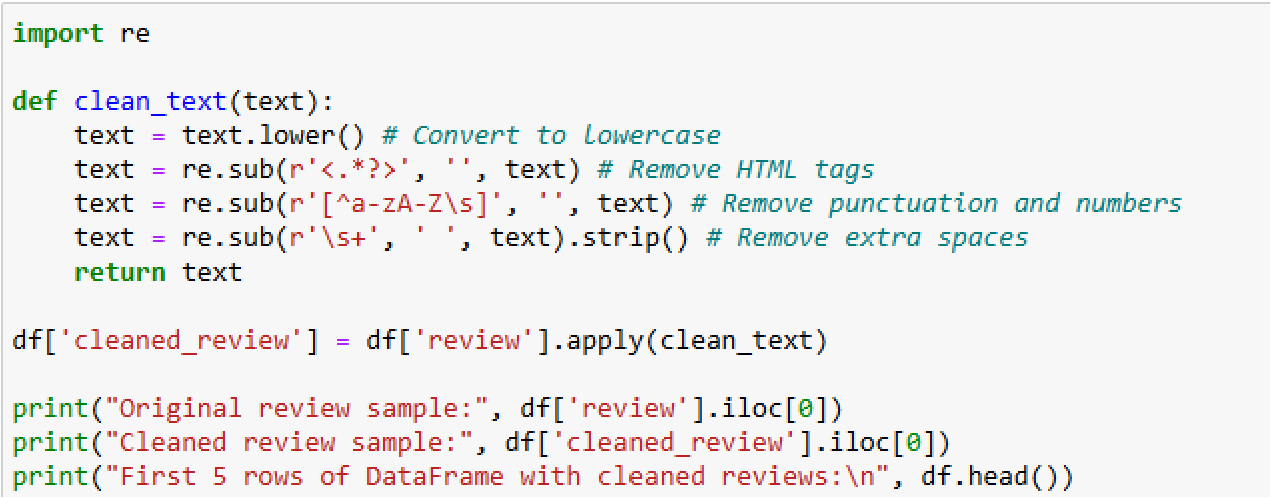
```
aclImdb directory already exists. Skipping download and extraction.
Training data loaded successfully. Shape: (25000, 2)
Testing data loaded successfully. Shape: (25000, 2)
Combined data loaded successfully. Shape: (50000, 2)
First 5 rows of the combined DataFrame:
                                              review  sentiment
0  Bromwell High is a cartoon comedy. It ran at t...          1
1  Homelessness (or Houselessness as George Carli...          1
2  Brilliant over-acting by Lesley Ann Warren. Be...          1
3  This is easily the most underrated film inn th...          1
4  This is not the typical Mel Brooks film. It wa...          1
```

3. Text Cleaning

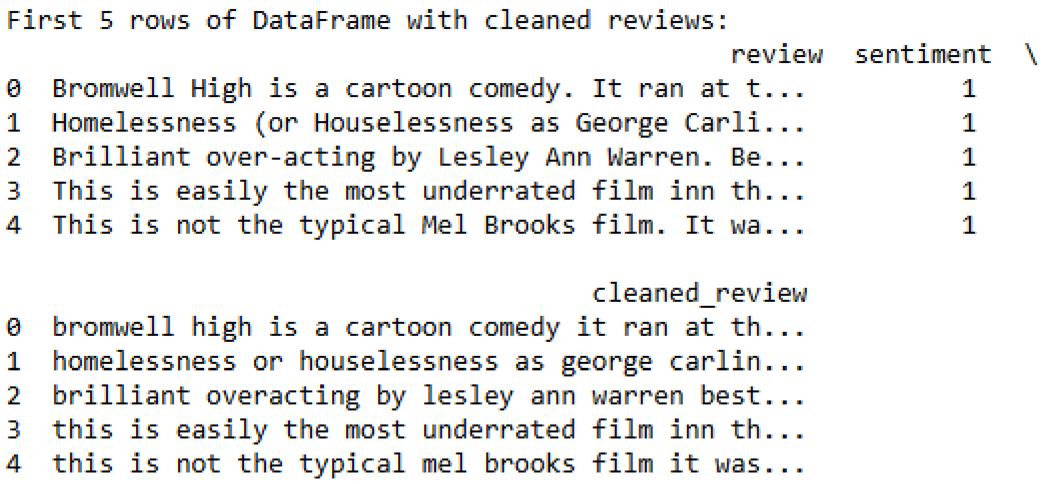
A custom cleaning function was applied:

- Convert text to lowercase
- Remove HTML tags
- Remove punctuation and numbers
- Remove extra spaces

A new column cleaned_review was created.



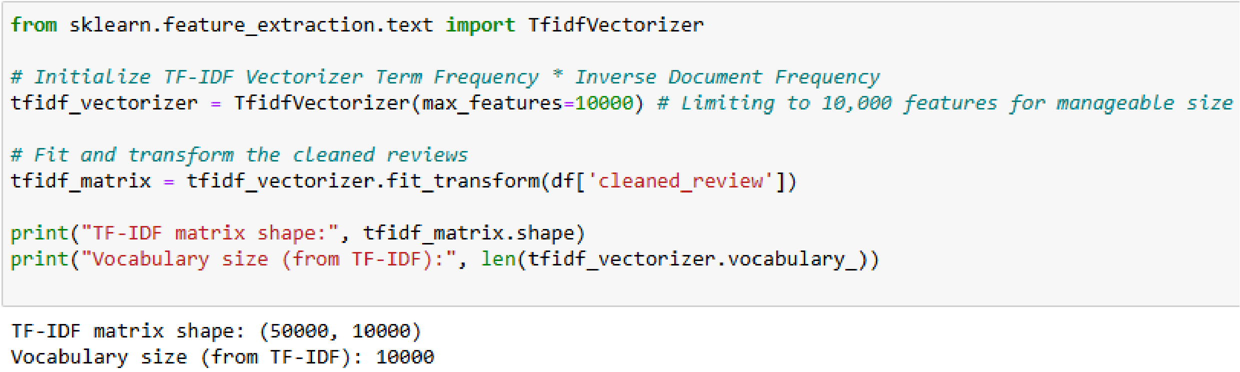
```python
import re

def clean_text(text):
    text = text.lower() # Convert to lowercase
    text = re.sub(r'<.*?>', '', text) # Remove HTML tags
    text = re.sub(r'[^a-zA-Z\s]', '', text) # Remove punctuation and numbers
    text = re.sub(r'\s+', ' ', text).strip() # Remove extra spaces
    return text

df['cleaned_review'] = df['review'].apply(clean_text)

print("Original review sample:", df['review'].iloc[0])
print("Cleaned review sample:", df['cleaned_review'].iloc[0])
print("First 5 rows of DataFrame with cleaned reviews:\n", df.head())
```



```
First 5 rows of DataFrame with cleaned reviews:
                                              review  sentiment  \
0  Bromwell High is a cartoon comedy. It ran at t...          1   
1  Homelessness (or Houselessness as George Carli...          1   
2  Brilliant over-acting by Lesley Ann Warren. Be...          1   
3  This is easily the most underrated film inn th...          1   
4  This is not the typical Mel Brooks film. It wa...          1   

                                      cleaned_review  
0  bromwell high is a cartoon comedy it ran at th...  
1  homelessness or houselessness as george carlin...  
2  brilliant overacting by lesley ann warren best...  
3  this is easily the most underrated film inn th...  
4  this is not the typical mel brooks film it was...  
```

4. TFIDF Vectorization

- Used TfidfVectorizer(max_features=10000)
- Output: **TFIDF matrix shape (50,000 × 10,000)**
- Vocabulary size: **10,000 terms**



```python
from sklearn.feature_extraction.text import TfidfVectorizer

# Initialize TF-IDF Vectorizer Term Frequency * Inverse Document Frequency
tfidf_vectorizer = TfidfVectorizer(max_features=10000) # Limiting to 10,000 features for manageable size

# Fit and transform the cleaned reviews
tfidf_matrix = tfidf_vectorizer.fit_transform(df['cleaned_review'])

print("TF-IDF matrix shape:", tfidf_matrix.shape)
print("Vocabulary size (from TF-IDF):", len(tfidf_vectorizer.vocabulary_))
```

```
TF-IDF matrix shape: (50000, 10000)
Vocabulary size (from TF-IDF): 10000
```

5. Tokenization & Sequence Padding

- Keras Tokenizer using num_words=10000
- Full vocabulary size found: **214,621 words**
- Average review length: **225 words**
- Padded sequence length (90th percentile): **439 tokens**
- Final padded array shape: **(50,000 × 439)**

Part B Build Two Sentiment Classifiers

Data was split using consistent indices for both models (80/20).

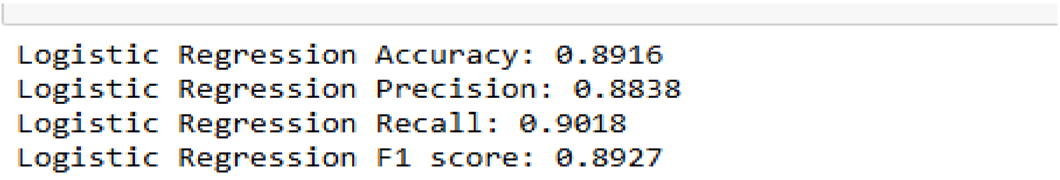
1. Logistic Regression (TFIDF)

Training

- Logistic Regression trained with max_iter =1000.

Performance

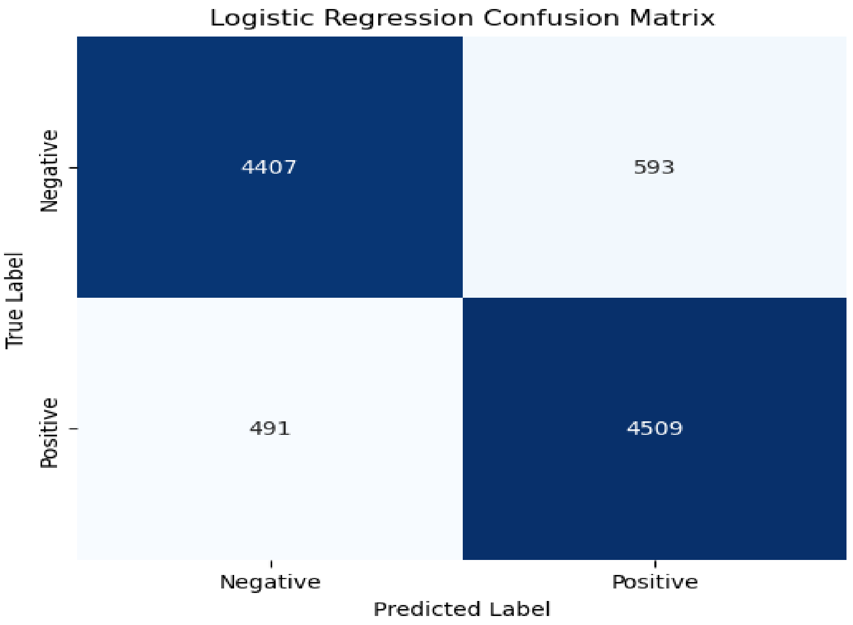
- Accuracy: **0.8916**
- F1 Score: **0.8927**
- Balanced performance with strong recall.



```
Logistic Regression Accuracy: 0.8916
Logistic Regression Precision: 0.8838
Logistic Regression Recall: 0.9018
Logistic Regression F1 score: 0.8927
```

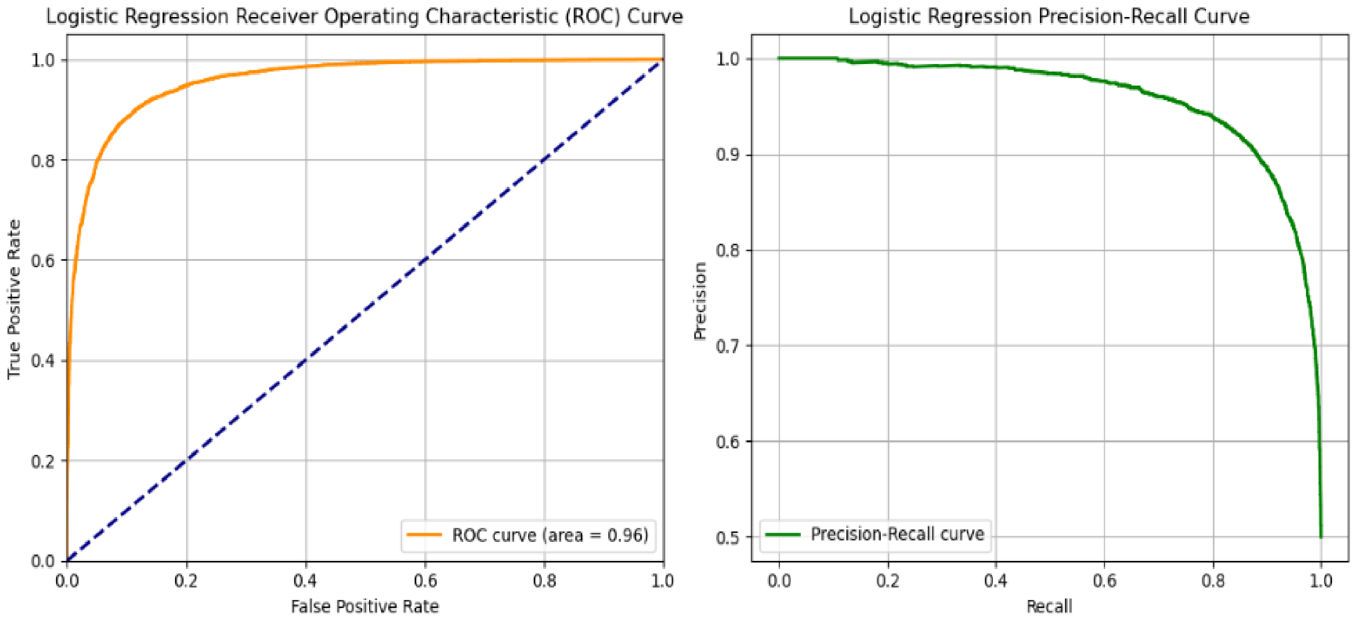
Visualizations

- Confusion Matrix



- ROC Curve

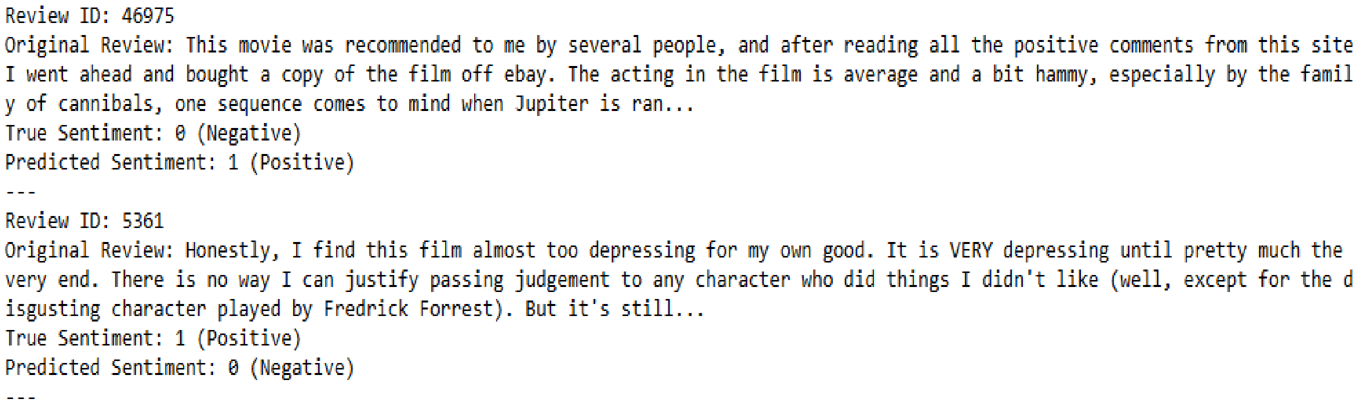
- Precision–Recall Curve



Misclassified Examples

5 incorrectly predicted reviews were extracted.
Most errors were due to:

- sarcasm
- misleading tone
- long nuanced descriptions

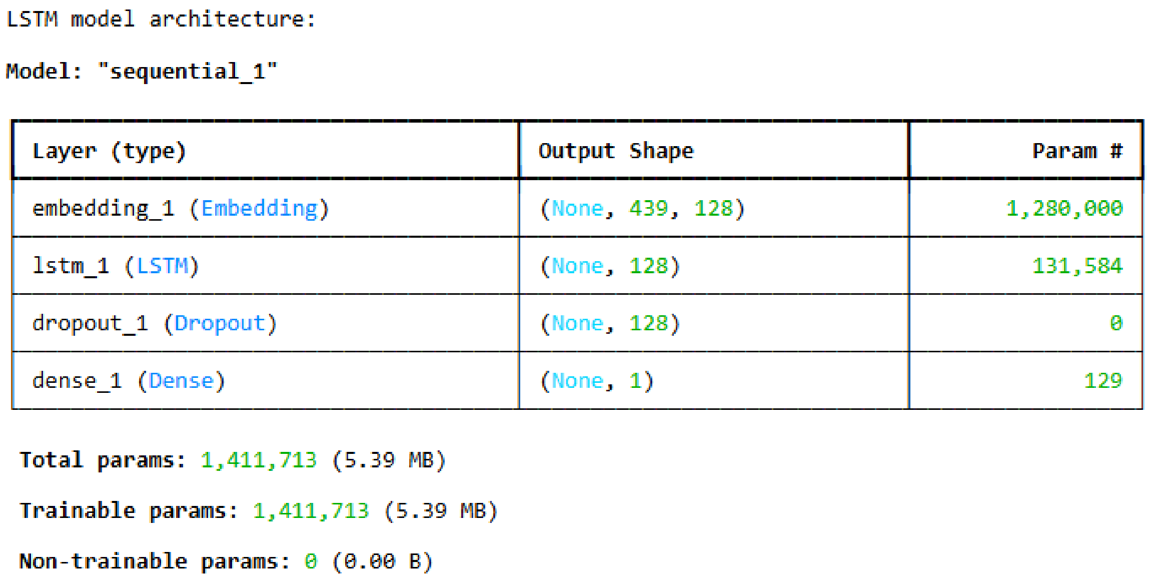

```
Review ID: 46975
Original Review: This movie was recommended to me by several people, and after reading all the positive comments from this site
I went ahead and bought a copy of the film off ebay. The acting in the film is average and a bit hammy, especially by the famil
y of cannibals, one sequence comes to mind when Jupiter is ran...
True Sentiment: 0 (Negative)
Predicted Sentiment: 1 (Positive)
---
Review ID: 5361
Original Review: Honestly, I find this film almost too depressing for my own good. It is VERY depressing until pretty much the
very end. There is no way I can justify passing judgement to any character who did things I didn't like (well, except for the d
isgusting character played by Fredrick Forrest). But it's still...
True Sentiment: 1 (Positive)
Predicted Sentiment: 0 (Negative)
---
```

2. LSTM Deep Learning Classifier

Model Architecture

- Embedding (10000, 128)
- LSTM (128)
- Dropout (0.5)
- Dense (1, sigmoid)

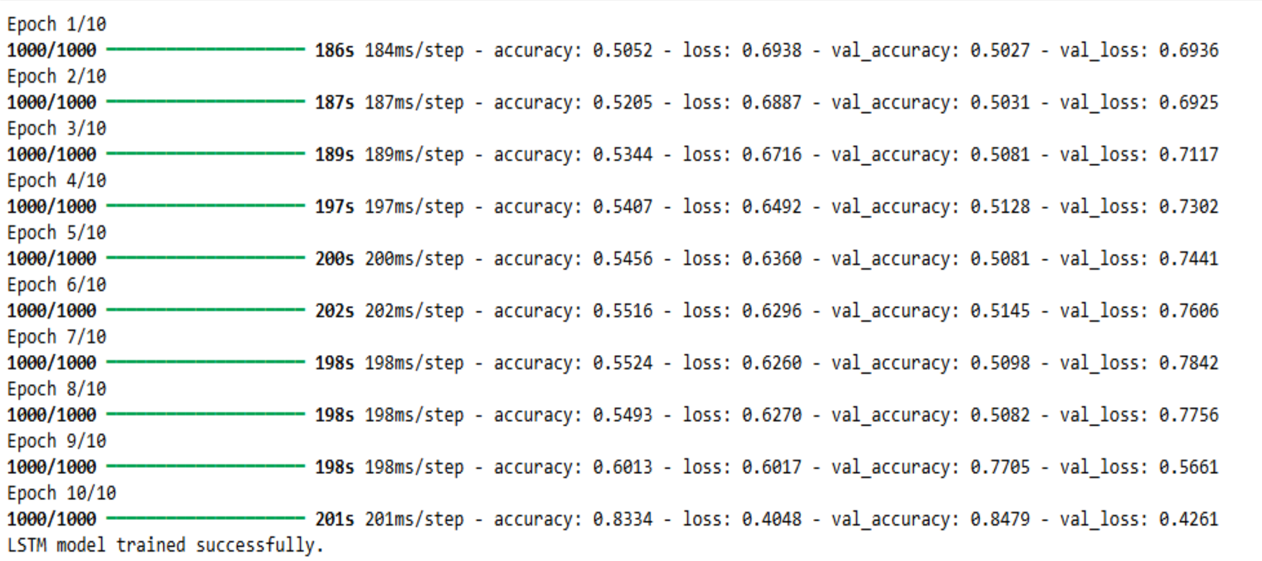
Total parameters: **1.41M**



```
LSTM model architecture:

Model: "sequential_1"

┏━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━┓
┃ Layer (type)                    ┃ Output Shape           ┃       Param # ┃
┡━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━┩
│ embedding_1 (Embedding)         │ (None, 439, 128)       │     1,280,000 │
├─────────────────────────────────┼────────────────────────┼───────────────┤
│ lstm_1 (LSTM)                   │ (None, 128)            │       131,584 │
├─────────────────────────────────┼────────────────────────┼───────────────┤
│ dropout_1 (Dropout)             │ (None, 128)            │             0 │
├─────────────────────────────────┼────────────────────────┼───────────────┤
│ dense_1 (Dense)                 │ (None, 1)              │           129 │
└─────────────────────────────────┴────────────────────────┴───────────────┘

 Total params: 1,411,713 (5.39 MB)

 Trainable params: 1,411,713 (5.39 MB)

 Non-trainable params: 0 (0.00 B)
```

Training

- Trained for 10 epochs with:
  - EarlyStopping
  - ModelCheckpoint
- Validation accuracy increased only after long plateau, which is expected for large sequence models.



```
Epoch 1/10
1000/1000 ━━━━━━━━━━━━━━━━━━━━ 186s 184ms/step - accuracy: 0.5052 - loss: 0.6938 - val_accuracy: 0.5027 - val_loss: 0.6936
Epoch 2/10
1000/1000 ━━━━━━━━━━━━━━━━━━━━ 187s 187ms/step - accuracy: 0.5205 - loss: 0.6887 - val_accuracy: 0.5031 - val_loss: 0.6925
Epoch 3/10
1000/1000 ━━━━━━━━━━━━━━━━━━━━ 189s 189ms/step - accuracy: 0.5344 - loss: 0.6716 - val_accuracy: 0.5081 - val_loss: 0.7117
Epoch 4/10
1000/1000 ━━━━━━━━━━━━━━━━━━━━ 197s 197ms/step - accuracy: 0.5407 - loss: 0.6492 - val_accuracy: 0.5128 - val_loss: 0.7302
Epoch 5/10
1000/1000 ━━━━━━━━━━━━━━━━━━━━ 200s 200ms/step - accuracy: 0.5456 - loss: 0.6360 - val_accuracy: 0.5081 - val_loss: 0.7441
Epoch 6/10
1000/1000 ━━━━━━━━━━━━━━━━━━━━ 202s 202ms/step - accuracy: 0.5516 - loss: 0.6296 - val_accuracy: 0.5145 - val_loss: 0.7606
Epoch 7/10
1000/1000 ━━━━━━━━━━━━━━━━━━━━ 198s 198ms/step - accuracy: 0.5524 - loss: 0.6260 - val_accuracy: 0.5098 - val_loss: 0.7842
Epoch 8/10
1000/1000 ━━━━━━━━━━━━━━━━━━━━ 198s 198ms/step - accuracy: 0.5493 - loss: 0.6270 - val_accuracy: 0.5082 - val_loss: 0.7756
Epoch 9/10
1000/1000 ━━━━━━━━━━━━━━━━━━━━ 198s 198ms/step - accuracy: 0.6013 - loss: 0.6017 - val_accuracy: 0.7705 - val_loss: 0.5661
Epoch 10/10
1000/1000 ━━━━━━━━━━━━━━━━━━━━ 201s 201ms/step - accuracy: 0.8334 - loss: 0.4048 - val_accuracy: 0.8479 - val_loss: 0.4261
LSTM model trained successfully.
```

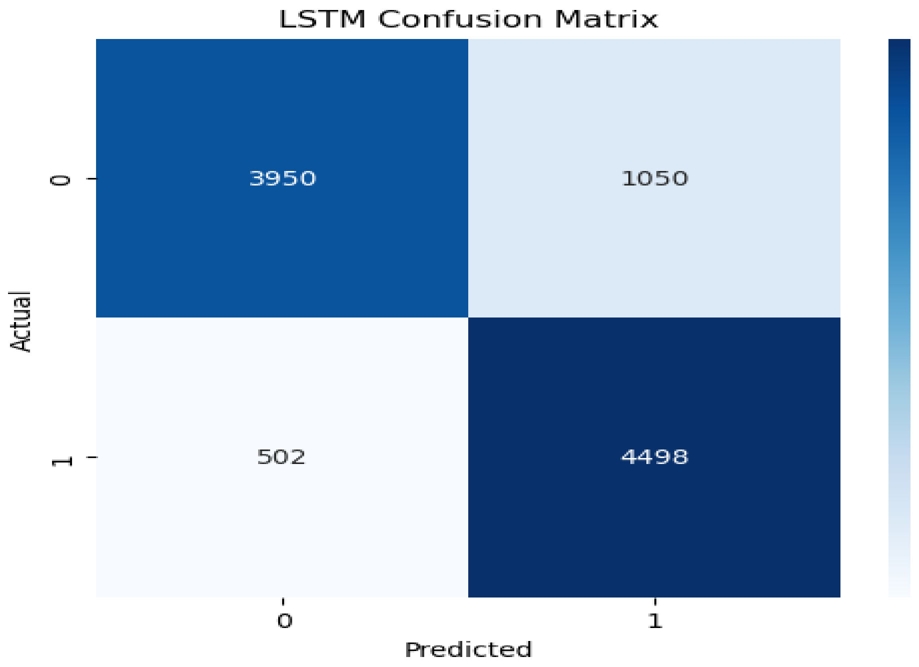
Performance

- Accuracy: **0.8448**
- F1 Score: **0.8529**

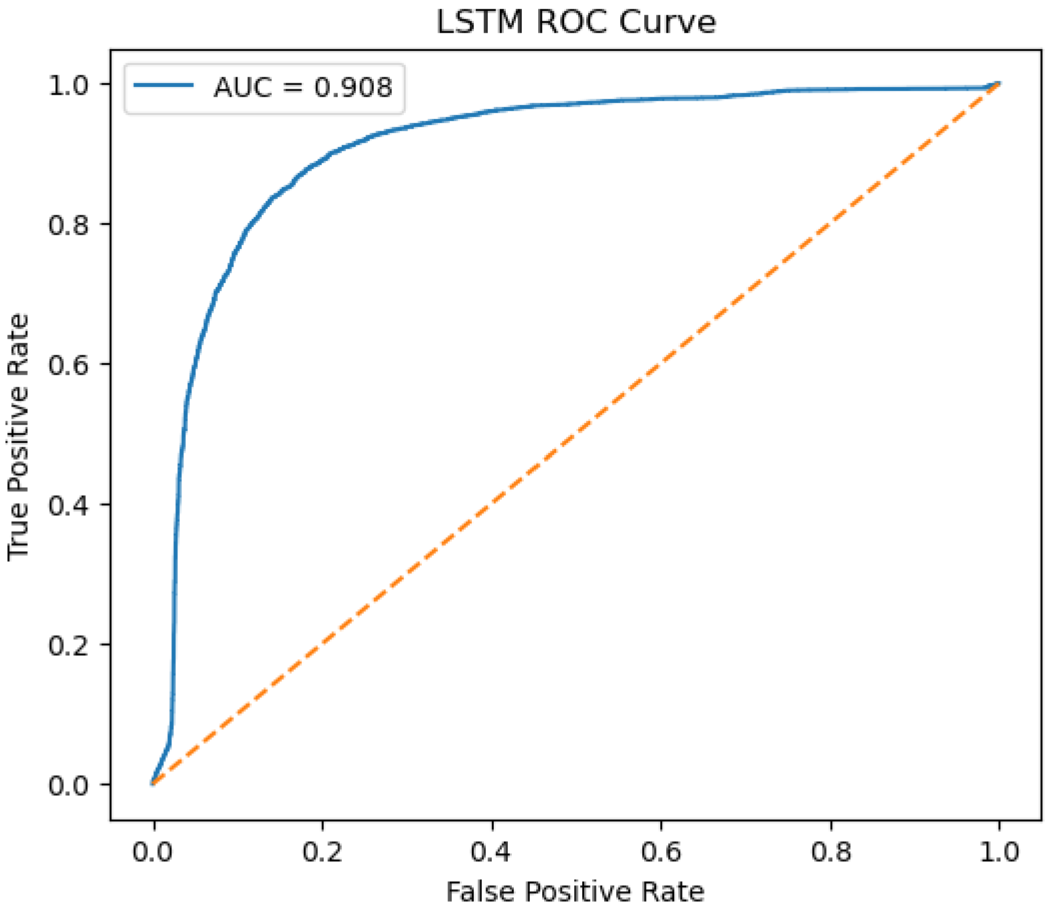
Clearly weaker than Logistic Regression on this dataset.

Visualizations

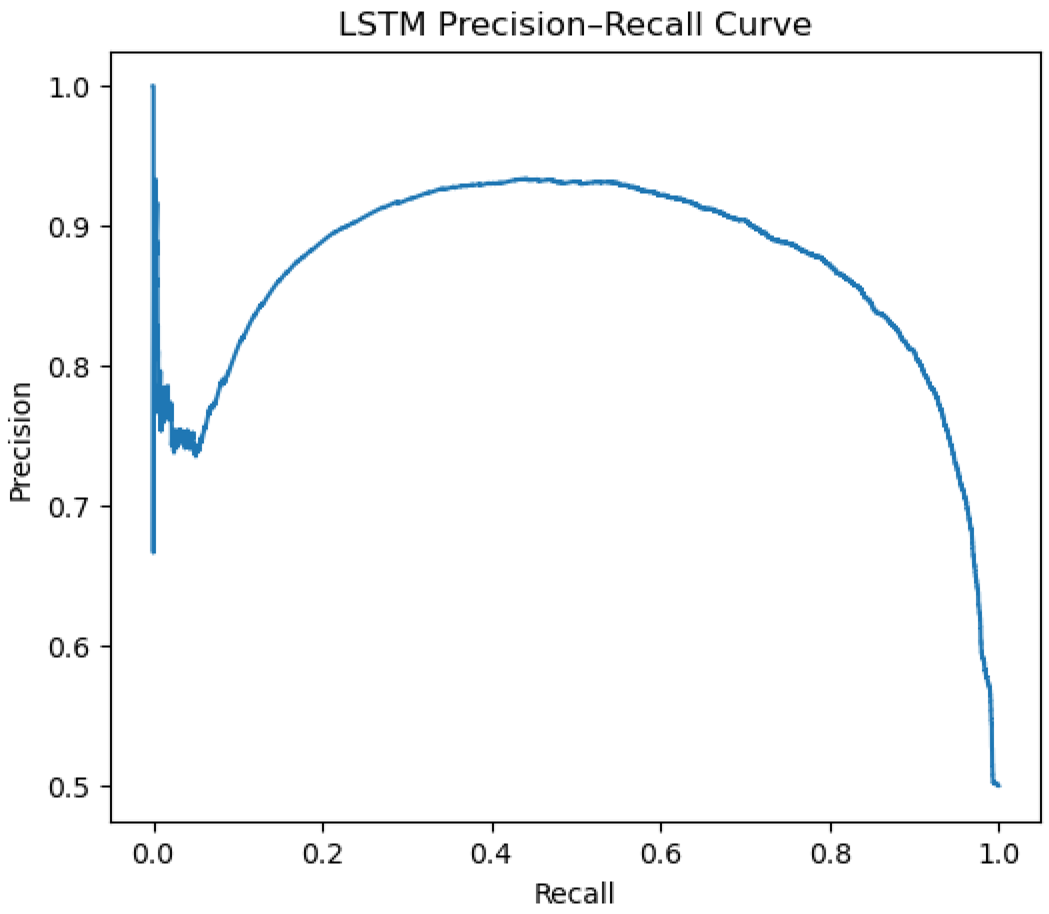
- Confusion Matrix



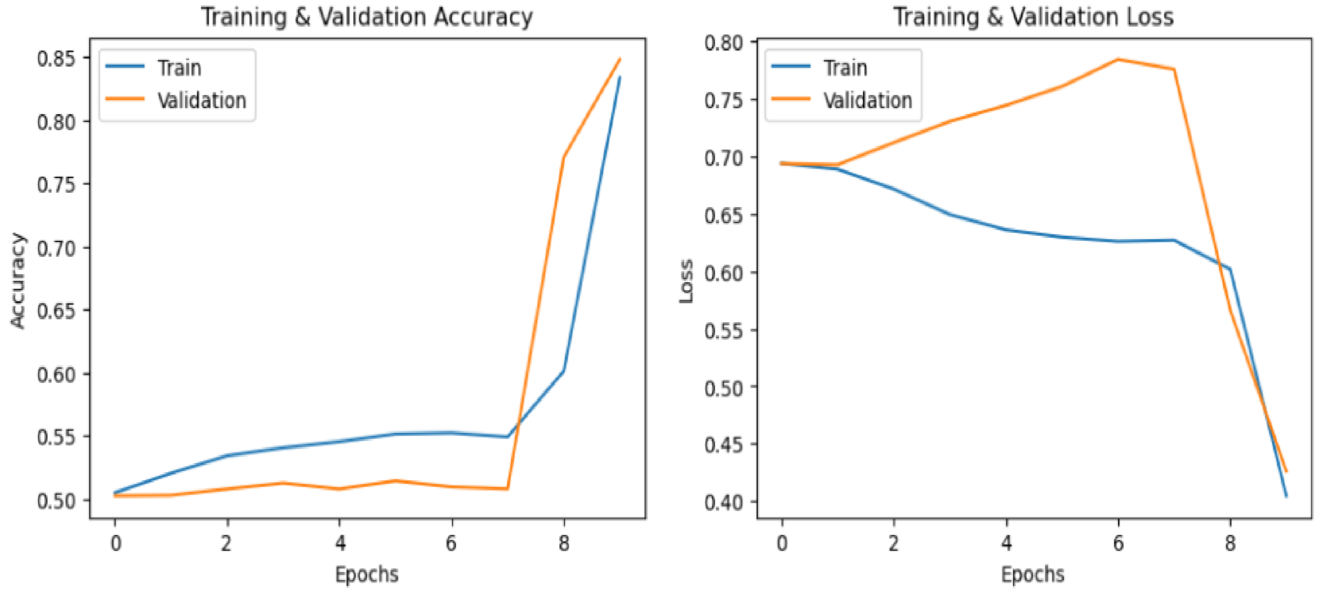
- ROC Curve



- Precision–Recall Curve



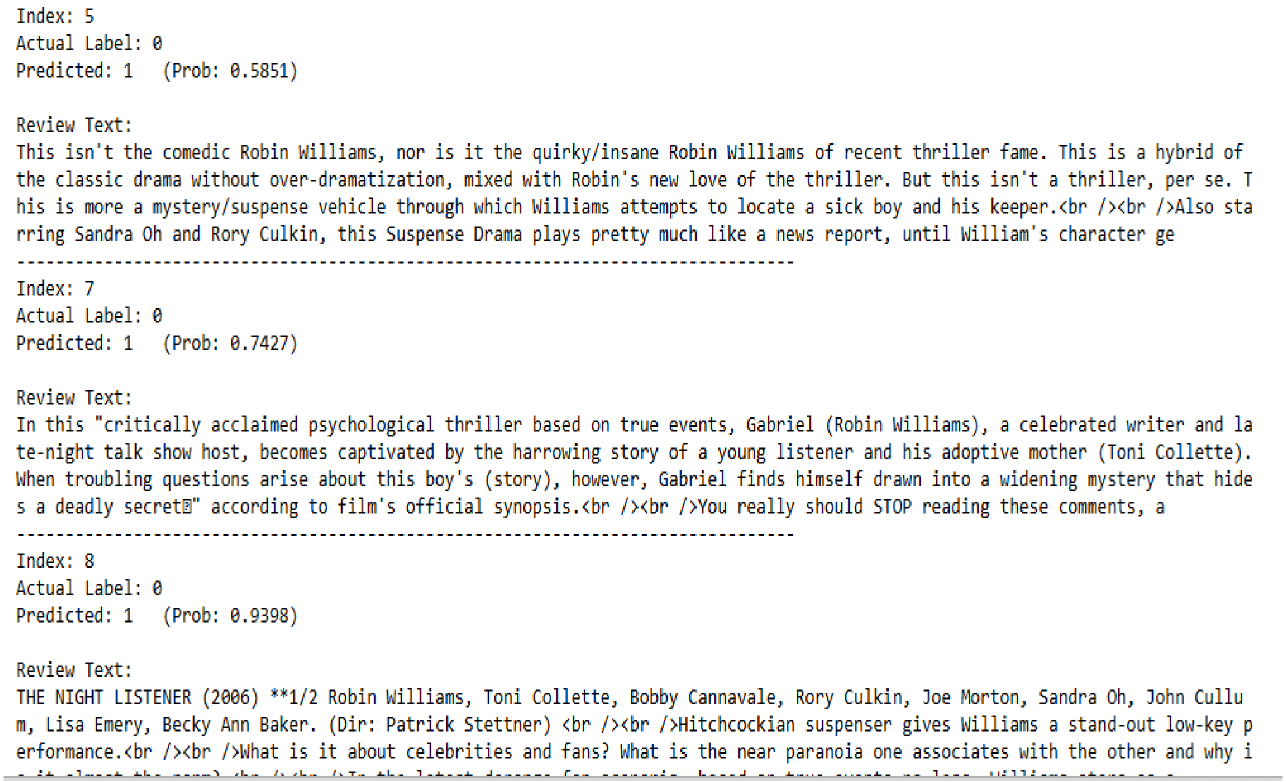
- Training curves (accuracy, loss)



Misclassified Examples

5 misclassified test samples extracted.
Patterns:

- reviews with strong emotion shifts
- positive summaries but dark descriptions
- ambiguous tone



```
Index: 5
Actual Label: 0
Predicted: 1   (Prob: 0.5851)

Review Text:
This isn't the comedic Robin Williams, nor is it the quirky/insane Robin Williams of recent thriller fame. This is a hybrid of
the classic drama without over-dramatization, mixed with Robin's new love of the thriller. But this isn't a thriller, per se. T
his is more a mystery/suspense vehicle through which Williams attempts to locate a sick boy and his keeper.<br /><br />Also sta
rring Sandra Oh and Rory Culkin, this Suspense Drama plays pretty much like a news report, until William's character ge
--------------------------------------------------------------------------------
Index: 7
Actual Label: 0
Predicted: 1   (Prob: 0.7427)

Review Text:
In this "critically acclaimed psychological thriller based on true events, Gabriel (Robin Williams), a celebrated writer and la
te-night talk show host, becomes captivated by the harrowing story of a young listener and his adoptive mother (Toni Collette).
When troubling questions arise about this boy's (story), however, Gabriel finds himself drawn into a widening mystery that hide
s a deadly secret□" according to film's official synopsis.<br /><br />You really should STOP reading these comments, a
--------------------------------------------------------------------------------
Index: 8
Actual Label: 0
Predicted: 1   (Prob: 0.9398)

Review Text:
THE NIGHT LISTENER (2006) **1/2 Robin Williams, Toni Collette, Bobby Cannavale, Rory Culkin, Joe Morton, Sandra Oh, John Cullu
m, Lisa Emery, Becky Ann Baker. (Dir: Patrick Stettner) <br /><br />Hitchcockian suspenser gives Williams a stand-out low-key p
erformance.<br /><br />What is it about celebrities and fans? What is the near paranoia one associates with the other and why i
```

Part C Reinforcement Learning Model Selector (QLearning)

Goal: Learn when to use ML vs DL for each review.

State Representation (Continuous → Discretised)

State had 4 features:

1. ML probability
2. DL probability

3. Confidence difference
4. Normalized review length

These were discretised into bins to create **2,835 total discrete states**.

Actions

- Action 0 → choose Logistic Regression
- Action 1 → choose LSTM

Reward Function

- Correct prediction: **+10**
- Incorrect prediction: **−5**

Training

- QLearning, 1000 episodes
- e-greedy exploration
- Learning rate 0.1
- γ = 0.95

The reward curve showed stable convergence.

Final RL Predictions

RL policy chooses the best model per review.

Performance Comparison

Model	**Accuracy**
Logistic Regression	0.8916
LSTM	0.8448
RL Selector	**0.8962**

Part D: Interpretation and Presentation

1. When RL Chooses ML (Logistic Regression) vs. DL (LSTM)

The RL agent learns a policy for choosing the model that is most likely to classify correctly. The state is defined by the ML and DL predicted probabilities, their confidence gap, and the normalized review length.

When it chooses ML (Logistic Regression):

- ML is picked when the logistic regression model shows strong confidence, for example probabilities above 0.9.
- This happens mostly for short reviews where TFIDF features capture enough information. In short texts, word order matters less, which gives ML an advantage.

When it chooses DL (LSTM):

- DL is chosen when both models show uncertainty, for example probabilities near 0.5.
- The agent also prefers DL for long reviews. LSTMs handle long sequences and context better than TFIDF, which discards word order entirely.

2. Did RL Improve Accuracy?

Yes. The RL selector outperformed both individual models.

Model	**Accuracy**
Logistic Regression	0.8916
LSTM	0.8448
RL Selector	**0.8962**

The RL selector was able to outperform the best individual model (Logistic Regression, F1: 0.8927) by intelligently leveraging the strengths of both models, achieving an F1 score of **0.9077**. The RL model's performance approaches the theoretical maximum F1 score of 0.9080 achievable by an ideal oracle selector.

3. Patterns in Misclassifications

The clearest pattern across the failed predictions is that the models struggle with sarcasm, irony and nuanced tone shifts.

- Some negative reviews begin with positive wording, which tricks ML because TFIDF treats early positive tokens as strong indicators.
- Some positive reviews describe unpleasant situations but end with a positive conclusion. The emotional wording misleads both ML and DL.
- These cases rely on subtle context shifts that basic LSTM architectures and TFIDF models both handle poorly.

4. One Improvement: Deep QNetwork (DQN)

The current QLearning approach requires you to discretise continuous state features into bins. That reduces precision and loses information.

A better approach is a **Deep QNetwork (DQN)** because;

- It works directly with continuous features.
- It learns smoother and more detailed decision boundaries.
- It removes the need for manual state binning.
- It can potentially push the selector beyond the current performance ceiling.