Part A: IMDb Movie Review Sentiment Analysis

1. Problem Statement

The goal of this project is to classify IMDb movie reviews into positive or negative sentiment categories. The objective is to use machine learning models to accurately predict the sentiment of a given review based on its text content.

2. Data Description

Dataset: IMDb movie reviews (50,000 rows).

Columns:

- review: Text of the review.

- sentiment: Label indicating positive or negative sentiment.

Additional Features:

- review_length: Number of characters in each review (created during EDA).

There are no null values, and only two unique sentiment classes: positive and negative.

3. EDA & Preprocessing

EDA Highlights:

- Review Length:
  - 95% of reviews are under 3,391 characters.
  - 10% of reviews are under 501 characters.
- Distribution:
  - The sentiment classes are balanced.
- Visuals:
  - Countplot of sentiment distribution.
  - Boxplot of review length by sentiment.

Preprocessing Steps:

- Converted reviews to lowercase.

- Removed punctuation, special characters, and stopwords.

- Applied tokenization and lemmatization.

- Created a cleaned review dataset for model training.

4. Models and Results

Models Implemented:

- Sentence transformers + Naive Bayes

- Sentence transformers + Logistic Regression

- Sentence transformers + SVM (Support Vector Machine)

- Sentence transformers + Random Forest

Evaluation Metrics:

- Accuracy

- Classification Report (Precision, Recall, F1-score)

Final Evaluation Results on Test Set (Hypertuned SVM):

1. Accuracy: 0.8285
2. Precision: 0.8231
3. Recall: 0.8368
4. F1 Score: 0.8299

Classification Report:

Label	Precision	Recall	F1-Score	Support
0	0.83	0.82	0.83	3125
1	0.82	0.84	0.83	3125
Accuracy			**0.83**	6250
Macro Avg	0.83	0.83	0.83	6250
Weighted Avg	0.83	0.83	0.83	6250

Key Results:

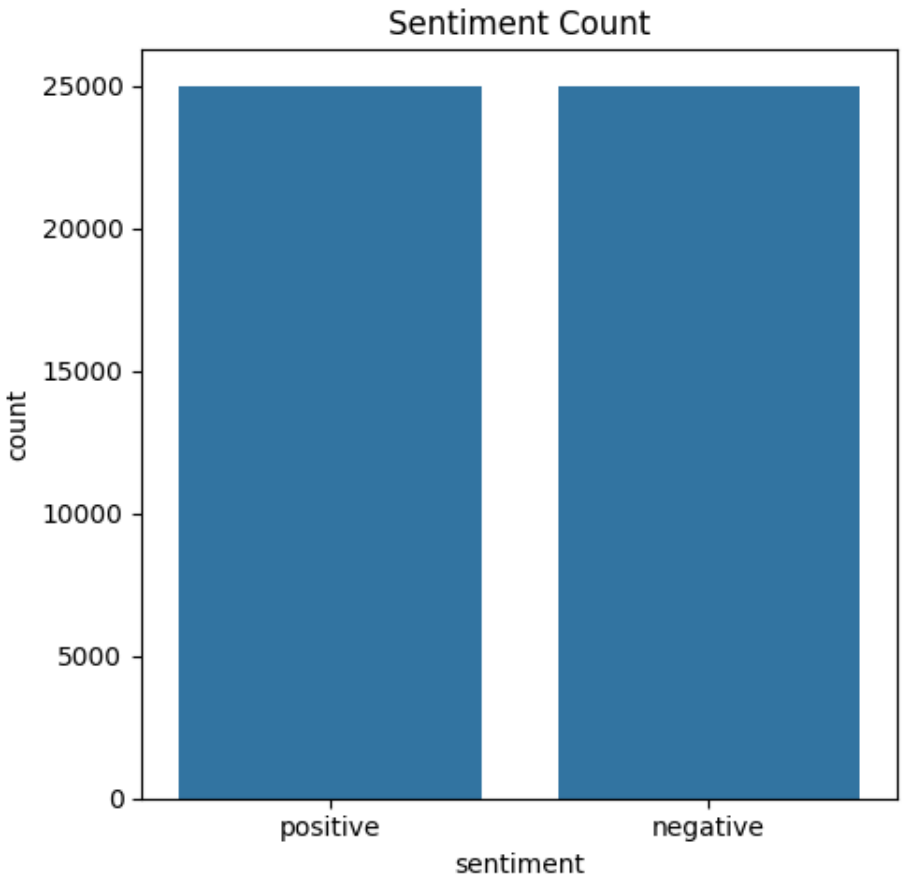
- Logistic Regression and SVM outperformed Naive Bayes in accuracy.

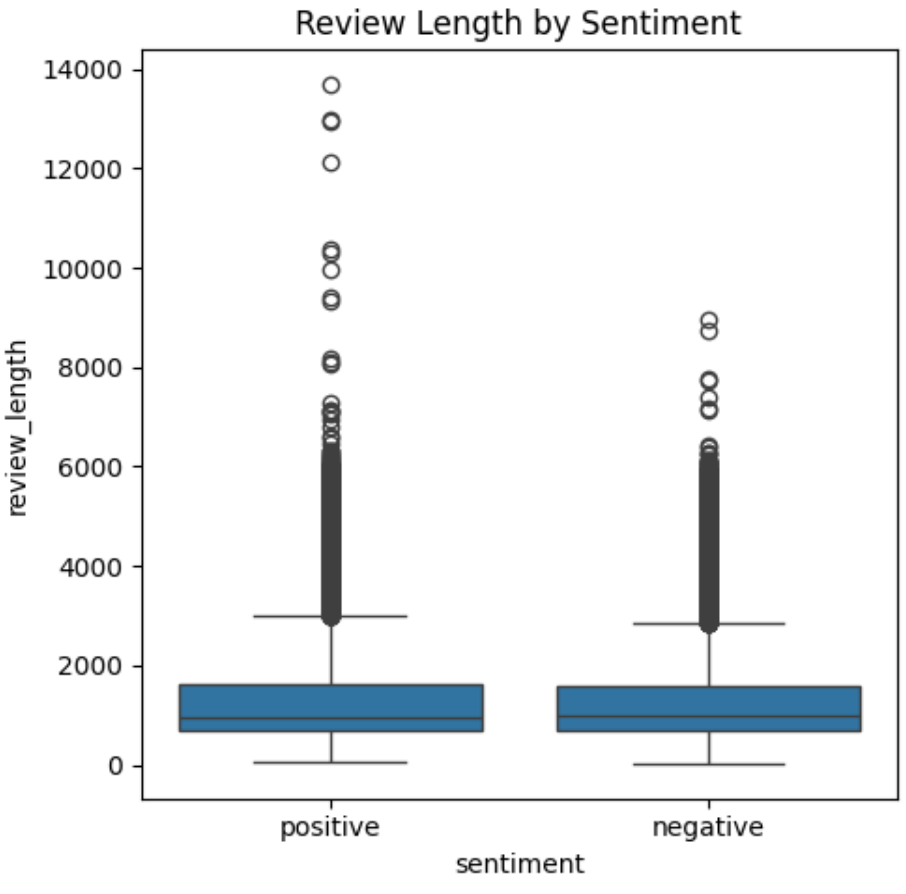
- SVM yielded the highest F1-score, indicating strong performance across both classes.

5. Visuals

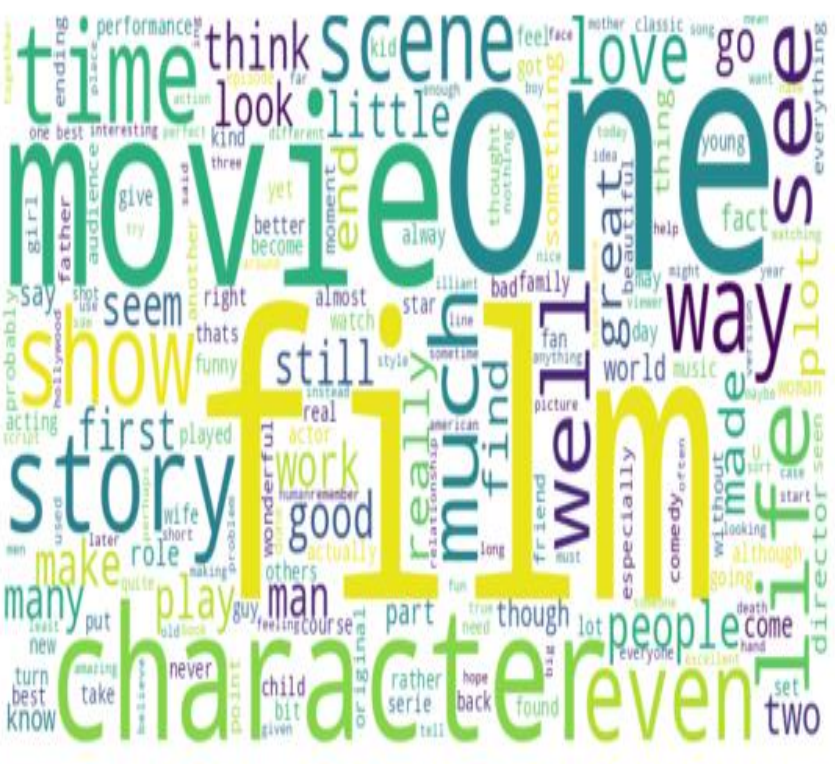
Visuals included:

- Count plot for sentiment distribution and boxplot for review length outliers.



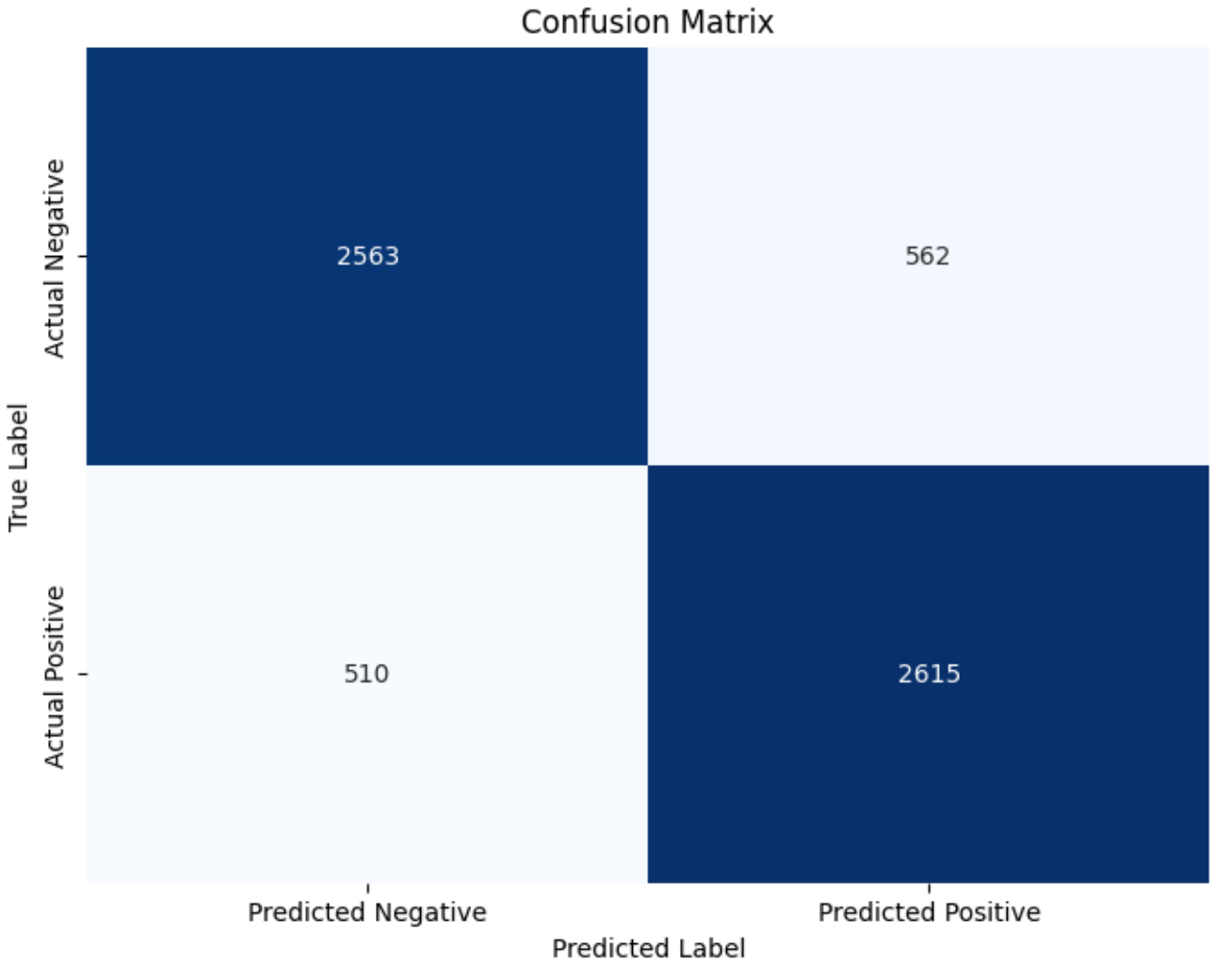


- Word clouds for positive and negative reviews.

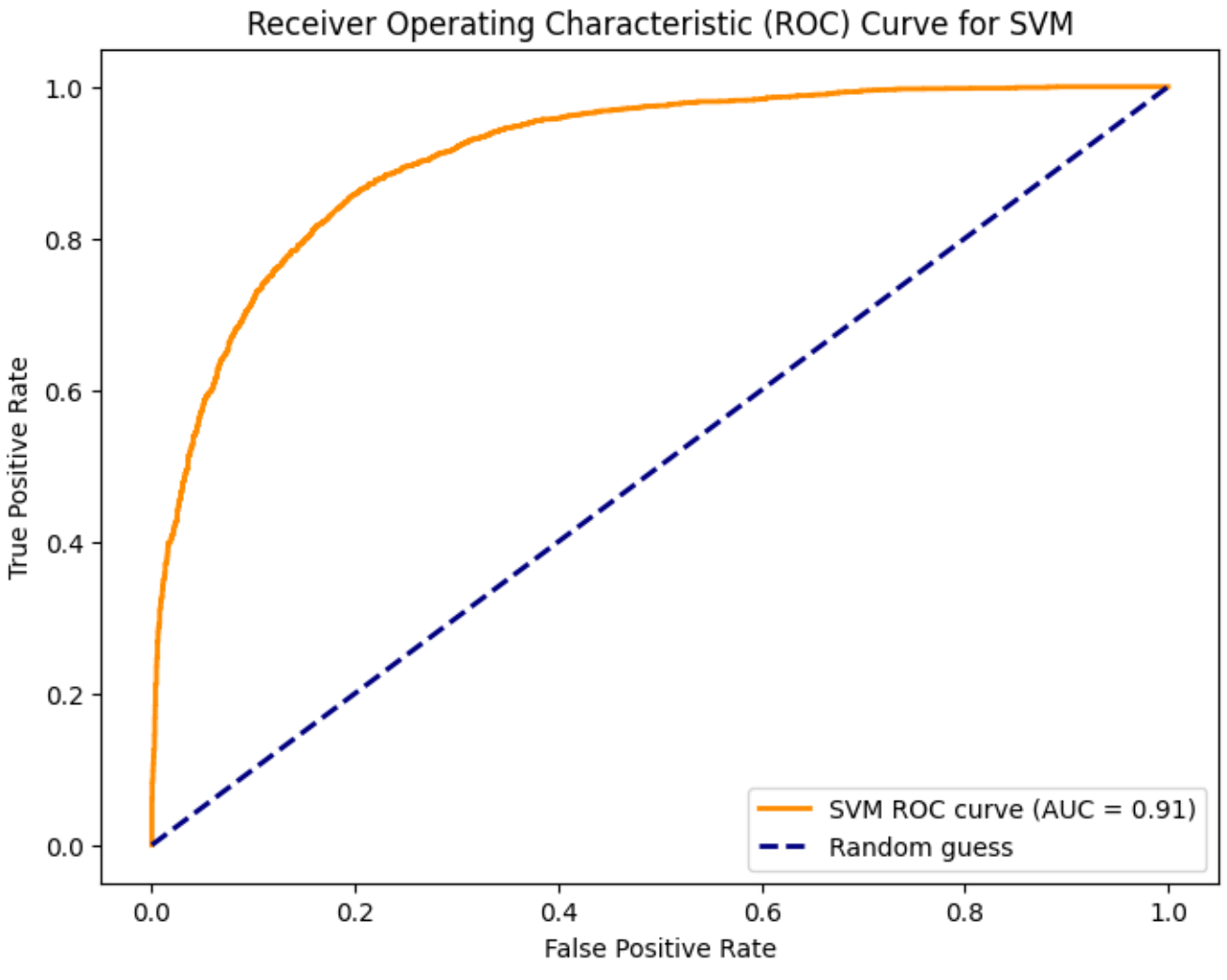


- Confusion matrix heatmaps for model evaluation.

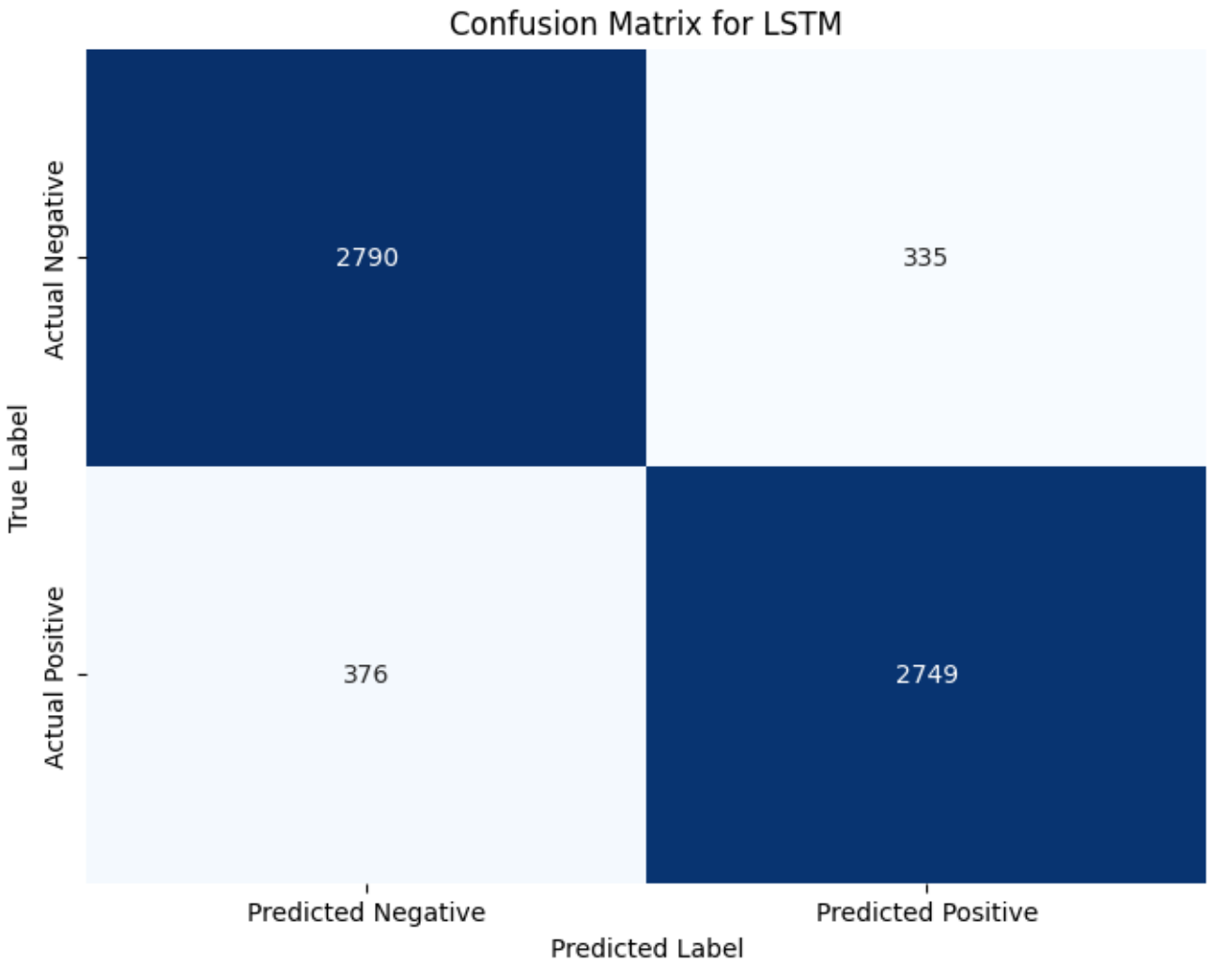
1. For hyper tuned SVM: -



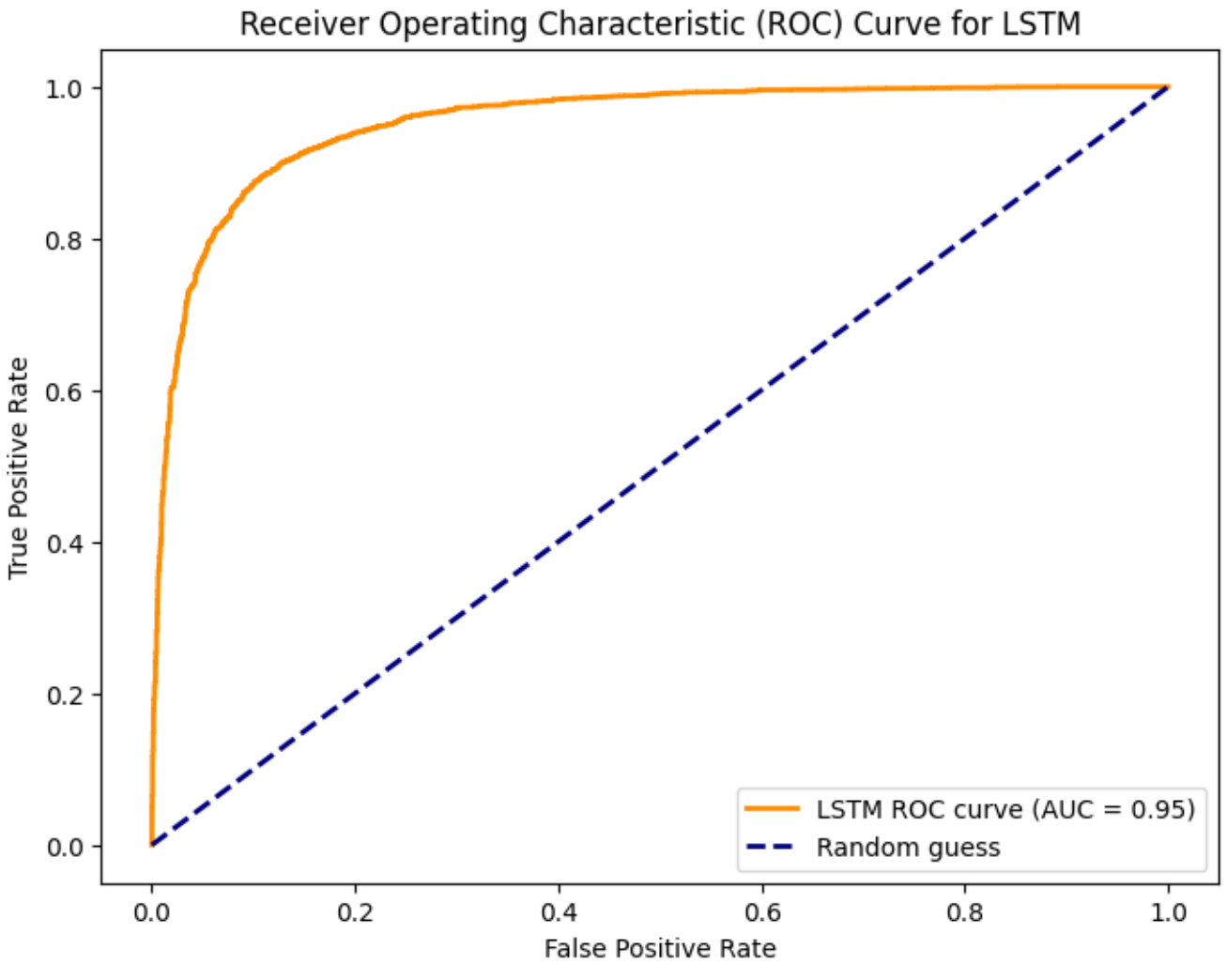
2. ROC-AUC curve for SVM Model: -



3. Confusion matrix for LSTM: -



4. ROC-AUC curve for LSTM



7. Insights & Conclusion:

- Frequent Words:
  - Positive Reviews: “great”, “amazing”, “loved”, “excellent”
  - Negative Reviews: “boring”, “worst”, “bad”, “waste”

- Sentiment Patterns:
  - Longer reviews tend to be positive, possibly due to more explanation.
  - Stop words removal and lemmatization significantly improved model clarity.
  - Sentence transformers proved more effective than TF-IDF in capturing review context.
- Implemented a function to input a new text and check its sentiment.
- Accuracy can further be increased by using better versions of sentence transformers.