Scenario V: Sentiment Analysis on Movie Reviews

1. Approach to Build ML Model

a. Dataset Description: The dataset provided has 12,500 entries each in labels of Positive and Negative and 50,000 entries in the Unsupervised label.

b. Selected Model: Due to the large number of entries in the dataset, I approached a simpler model so that it would be trainable with limited computing resources. Since we have labeled data, I constructed a model to perform binary classification using Logistic Regression since the data is already labeled.

c. Using NLTK, it was relatively faster to preprocess the text by tokenization, stopword removal, and lemmatization.

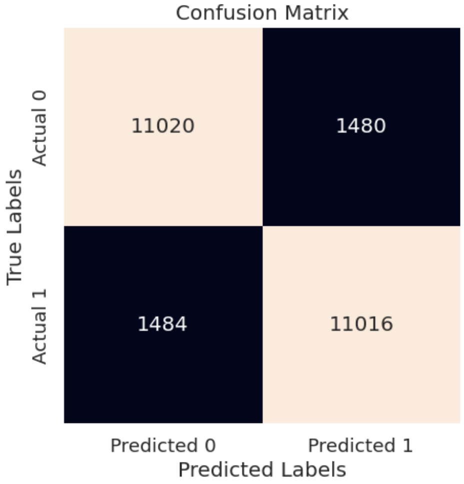
d. From sklearn.feature_extraction.text, I used TfidfVectorizer to vectorize the text data so that we have numerical data to construct a model on.

e. Ultimately, the model was trained within a few seconds.

2. Model Performance

a. The Accuracy Score for the generated model is 0.88144. Hence, about 88% of the predicted results were correct.

b. From this confusion matrix:



	Predicted 0	Predicted 1
Actual 0	11020	1480
Actual 1	1484	11016

We can observe TP, TN, FP and, FN.

c. Classification Report shows that the accuracy for both pos and neg predictions are equal which brightens the fact that our dataset was well-balanced.

	precision	recall	f1-score	support
neg	0.88	0.88	0.88	12500
pos	0.88	0.88	0.88	12500
accuracy			0.88	25000
macro avg	0.88	0.88	0.88	25000
weighted avg	0.88	0.88	0.88	25000

3. Potential Improvements

a. As you can observe, 88% is a good score for this task using Logistic Regression since the accuracy score for this model for a good dataset lies between 80% - 90%.

b. Since we have room for improvement, we can use advanced NLP models like BERT to perform binary classification on text data with an accuracy score over 90% - 95%.

c. Although nimble, Logistic Regression has limited fine-tuning capability, hence switching to another model would be a better idea.