BSDSF22A025

Lubaba Khalid

Assignment 7

Report

Sentiment Analysis using Logistic Regression & LSTM

Dataset: IMDB Movie Reviews Dataset

Description: The dataset contains 50,000 movie reviews, labeled as either positive (1) or negative (0).

Split: 25,000 reviews for training and 25,000 for testing.

Preprocessing:

Reviews are pre-tokenized as sequences of integers.

Maximum vocabulary size limited to 10,000 most frequent words.

Padding applied to ensure uniform input length for LSTM model.

Key Steps and Methods
Decoding Reviews:

The integer-encoded reviews were converted back to plain English text for TF-IDF vectorization.

Model 1 : Logistic Regression (TF-IDF):

Reviews were transformed into numerical vectors using TfidfVectorizer with a max feature size of 6000.

A Logistic Regression classifier was trained on this representation.

Evaluated using accuracy, precision, recall, and F1-score.

Model 2 : LSTM (Deep Learning):

Reviews were padded to a max length of 250.

An LSTM model with an embedding layer was trained for 3 epochs.

Accuracy on the test set was used for performance comparison.

Training vs Validation accuracy visualized.

Comparison of Results

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression (TF-IDF)	~0.88	~0.87	~0.89	~0.88
LSTM (Deep Learning)	~0.86	N/A	N/A	N/A

Logistic Regression performed slightly better in this configuration.

LSTM could improve with more training epochs or hyperparameter tuning.

Final Thoughts & Recommendations

TF-IDF + Logistic Regression provides a strong and fast baseline for text classification tasks.

LSTM models capture sequential context and semantics better but require more computation and tuning.