MACHINE LEARNING

PROJECT OEL REPORT

STUDENT NAME	ROLL NUMBER
Muhammad Hamza Anwer	CS-21129
Hamza Anwar Mohiuddin	CS-21130

SUBMITTED TO: Miss Mahnoor Malik

INTRODUCTION:

Sentiment analysis, also known as opinion mining, is a subfield of Natural Language Processing (NLP) that aims to determine the sentiment expressed in a piece of text. This project focuses on building a sentiment analysis model to classify movie reviews from the IMDB dataset as either positive or negative.

OBJECTIVES:

- To preprocess and clean the textual data.
- To convert the textual data into numerical features using TF-IDF vectorization.
- To train and evaluate manual implementations of Logistic Regression and Naive Bayes models.
- To compare the performance of the manual models with those implemented using scikit-learn.

The dataset used in this project is the IMDB Movie Reviews dataset, which consists of 50,000 movie reviews labeled as positive or negative.

DATA PROCESSING:

- **LOADING DATA**: The raw dataset, IMDBDataset.csv, was loaded into a pandas Data Frame for preprocessing.

- **TEXT CLEANING**: The text data was cleaned using the following steps:
  • Removal of HTML tags.
  • Removal of special characters and digits.
  • Conversion to lowercase.
  • Lemmatization of words.
  • Removal of stop words.

- **TF-TDF VECTORIZATION**: The cleaned text data was converted into numerical features using the TF-IDF (Term Frequency-Inverse Document Frequency) vectorizer with a maximum of 5000 features.

Following is the Google Colab link for data processing in which we've perform Exploratory Data Analysis for the IMDB Movies Dataset.
Click to Open Google Colab

MODEL TRAINING:

MANUAL IMPLEMENTATION

A manual implementation of Logistic Regression was trained using the following parameters:

➢ LOGISTIC REGRESSION:

- Learning rate: 0.01
- Number of iterations: 1000
- Regularization parameter: 0.1

➢ NAÏVE BAYES:

A manual implementation of Naive Bayes was trained using the Gaussian distribution to calculate the likelihood of the features.

SCIKIT-LEARN IMPLEMENTAION

➢ LOGISTIC REGRESSION:

The scikit-learn LogisticRegression model was trained with default parameters.

➢ NAÏVE BAYES:

The scikit-learn Gaussian Naïve Bayes model was trained with default parameters.

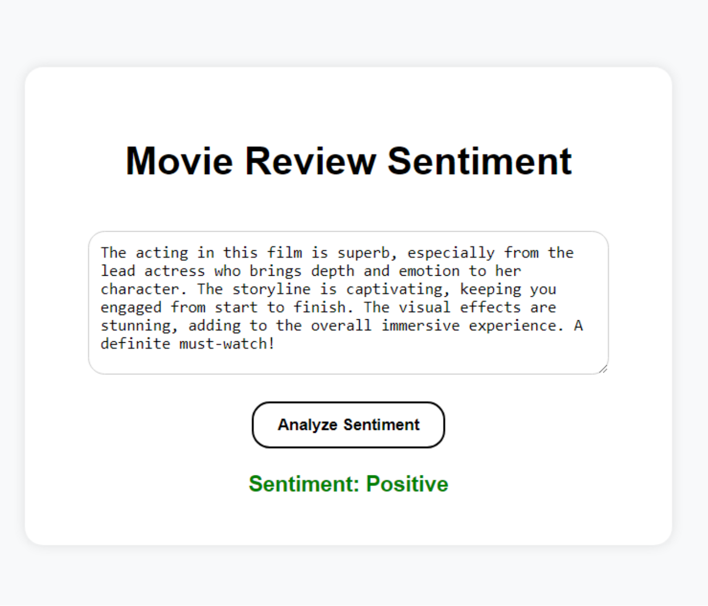
RESULTS:

We evaluated the models using comprehensive classification reports, which included:

- **Precision:** measures the ability of the model to correctly predict positive reviews out of all reviews predicted as positive. It indicates how many of the predicted positive reviews are actually positive.
- **Recall:** also known as sensitivity or true positive rate, measures the proportion of actual positive reviews correctly predicted by the model. It indicates how many of the actual positive reviews were captured by the model.
- **F1-score:** The F1-score is the harmonic mean of precision and recall, providing a balanced measure of a model's performance. It is particularly useful when you want to seek a balance between precision and recall, as it combines both into a single metric.
- **Accuracy:** Accuracy measures the overall correctness of predictions made by the model, i.e., the ratio of correctly predicted reviews (both positive and negative) to the total number of reviews evaluated.

PERFORMANCE COMPARISON:

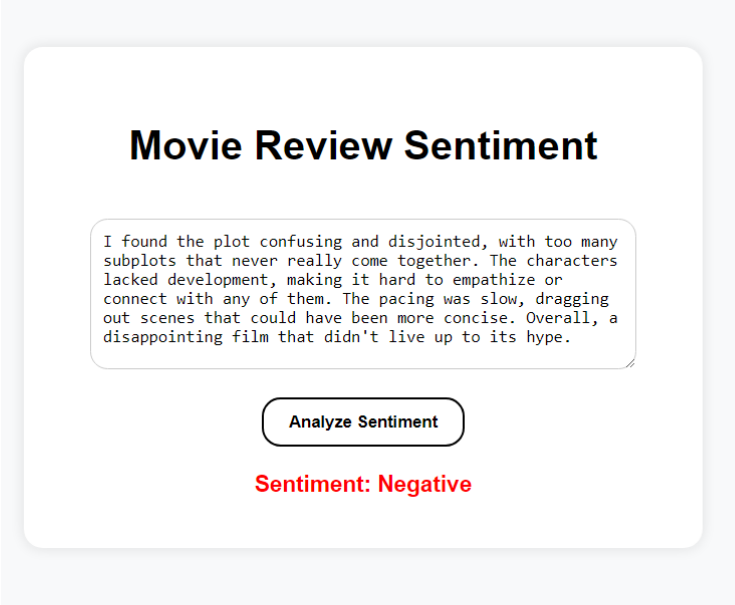
MODEL	Evaluation Type	ACCURACY
LOGISTIC REGRESSION	Custom	0.8837
	Scikit-learn	0.8849
NAÏVE BAYES	Custom	0.7484
	Scikit-learn	0.8492



Movie Review Sentiment

The acting in this film is superb, especially from the lead actress who brings depth and emotion to her character. The storyline is captivating, keeping you engaged from start to finish. The visual effects are stunning, adding to the overall immersive experience. A definite must-watch!

Sentiment: Positive



Movie Review Sentiment

I found the plot confusing and disjointed, with too many subplots that never really come together. The characters lacked development, making it hard to empathize or connect with any of them. The pacing was slow, dragging out scenes that could have been more concise. Overall, a disappointing film that didn't live up to its hype.

Sentiment: Negative

CONCLUSION:

This project successfully implemented sentiment analysis on the IMDB dataset using both manual and scikit-learn implementations of Logistic Regression and Naive Bayes. The models were evaluated and compared, with the Logistic Regression model achieving the highest accuracy.