# Sentiment Analysis Using Machine Learning/Artificial Intelligence

## 1. Introduction

This project focuses on sentiment analysis using machine learning techniques. The goal is to classify movie reviews as either positive or negative. The dataset used is the IMDB dataset, which contains labeled movie reviews.

## 2. Approach Used

The sentiment analysis was implemented using the following steps:

### 2.1 Data Preprocessing

* Loaded the dataset and handled potential issues like missing values.
* Mapped sentiment labels to numerical values (1 for positive, 0 for negative).
* Cleaned the text data by:

1. Removing HTML tags.
2. Removing non-alphabetic characters.
3. Converting text to lowercase.
4. Tokenizing words.
5. Removing stopwords.
6. Lemmatizing words.

### 2.2 Feature Extraction

* Applied **TF-IDF (Term Frequency-Inverse Document Frequency)** vectorization to convert text into numerical form.
* Limited features to 5000 most important words.

### 2.3 Model Training

* Split the dataset into training (80%) and testing (20%) sets.
* Trained a **Logistic Regression** model to classify reviews.

### 2.4 Model Evaluation

Evaluated the trained model using:

1. **Accuracy**
2. **Precision**
3. **Recall**
4. **F1-score**

### 2.5 Sentiment Prediction

Created a function to take user input, preprocess it, and predict the sentiment using the trained model.

## 3. Challenges Faced

* Handling missing or corrupted data.
* Dealing with stopwords and lemmatization to improve text clarity.
* Optimizing feature extraction for better accuracy while reducing computation time.

## 4. Model Performance

**Accuracy:** 0.5278

**Precision:** 0.5085

**Recall:** 0.8571

**F1-score:** 0.6383

## 5. Improvements

* Using more advanced text embedding techniques such as Word2Vec or BERT.
* Fine-tuning hyperparameters to improve model accuracy.

## 6. Instructions for Running the Code

1. Install dependencies and Download NLTK dataset:

**pip install pandas numpy nltk scikit-learn**

**import nltk**

**nltk.download('punkt')**

**nltk.download('stopwords')**

**nltk.download('wordnet')**

1. Run the Python script or Jupyter Notebook.
2. Input a review to test the sentiment prediction.