**Text Sentiment Analysis using Logistic Regression**

**Name:** KHANSA AHMAD QURESHI

**Date:** 16th April 2025

**Dataset Used:** IMDB Dataset.csv from Kaggle

# Objective

This project focuses on building a sentiment analysis model using the IMDB movie review dataset. The objective was to classify reviews as either positive or negative using a Logistic Regression classifier.

## Methodology

The dataset was preprocessed by performing the following steps:

* Tokenization of the text into individual words
* Removal of stopwords to eliminate non-informative words
* Lemmatization for text normalization

After preprocessing, the text data was converted into numerical format using TF-IDF vectorization. The resulting features were then used to train a Logistic Regression classifier.

## Model Training and Evaluation

We used a Logistic Regression model with a maximum iteration limit of 1000. The model was trained on 80% of the dataset and tested on the remaining 20%. Evaluation metrics included accuracy, precision, recall, F1-score, and a confusion matrix. The model achieved high performance in accurately predicting the sentiment of movie reviews.

## Results

The Logistic Regression model demonstrated the following outcomes:

* High accuracy on the test dataset
* Balanced precision and recall for both classes
* A clear separation of positive and negative reviews as shown in the confusion matrix

## Challenges and Solutions

During the project, several challenges were encountered:

* Handling noisy and unstructured text data
* Ensuring accurate lemmatization and removal of stopwords
* Avoiding overfitting during model training

These challenges were addressed through proper preprocessing, tuning the Logistic Regression model, and visualizing the confusion matrix to interpret the classification results effectively.

## Conclusion

The sentiment analysis model using Logistic Regression proved to be effective for binary classification of movie reviews. The project helped in understanding essential NLP tasks such as text preprocessing, feature engineering, and model evaluation.