ARTIFICIAL INTELLIGENCE & MACHINE LEARNING

PROJECT\_B

SEC : 5

NAMES :

1. VENKAT MANOJ ( 2320030111)
2. BALAJI (2320030016)
3. ADITYA VINAY (2320030179)

PROBLEM STATEMENT :

Title: Sentiment Analysis

Description:

In the age of digital media, online reviews have become a significant source of information for consumers from which they can either judge a movie or a online product. However, manually analyzing and summarizing the sentiment of these reviews can be time-consuming and subjective. To address this challenge, the goal is to develop an automated system that can classify the sentiment of reviews as positive or negative.

Data set:

**1. Amazon Reviews Data-set**

Description: Includes product reviews from Amazon, categorized into different sentiment labels. It has a large number of reviews, including movie-related ones.

Source: Amazon Customer Reviews Data-set

Format: Reviews with ratings and metadata. You can filter for movie-related reviews.

1. **IMDb Movie Reviews Data-set**

Description: Contains 50,000 movie reviews (25,000 positive and 25,000 negative) from IMDb. This data-set is widely used for sentiment analysis tasks.

Source: [IMDb Movie Reviews Dataset](https://ai.stanford.edu/~amaas/data/sentiment/" \t "_new)

Format: Reviews in text format with labels (positive or negative).

**3. The Movie Database (Tm-db) API**

Description: Use TM-Db’s API to gather movie reviews and ratings. You’ll need to preprocess the data to label sentiments.

Source: Tm-db API

Format: Reviews and metadata, requiring some preprocessing for sentiment analysis.

Algorithm:

**Naive Bayes Classifier**

Overview: Naive Bayes is a probabilistic classifier based on Bayes' theorem, assuming that the presence of a feature in a document is independent of the presence of any other feature. It works well for text classification tasks due to its simplicity and efficiency.

Steps to Implement Naive Bayes for Sentiment Analysis:

**1.Text Preprocessing:**

Tokenization: Split text into individual words or tokens.

Stop Words Removal: Remove common words that don’t contribute much to the sentiment (e.g., “and”, “the”).

Stemming/Lemmatization: Reduce words to their base or root form (e.g., “running” to “run”).

**Feature Extraction:**

Bag of Words (BoW): Represent text data as a matrix of token counts.

Term Frequency-Inverse Document Frequency (TF-IDF): Scale the token counts by their importance in the dataset.

**Model Training:**

Train-Test Split: Split the data into training and testing sets.

Training: Use the training set to fit the Naive Bayes model.

**Model Evaluation:**

Prediction: Predict the sentiment of text in the test set.

Metrics: Evaluate the model using accuracy, precision, recall, and F1-score.

EXPECTED OUTCOME:

The project will result in a sentiment analysis model that accurately classifies any type of product or movie reviews as either positive or negative. You’ll achieve this using a Naive Bayes classifier, which will be evaluated for performance with metrics like accuracy, precision, recall, and F1-score. The model will predict sentiments for new reviews, helping to gauge overall opinions about movies. Additionally, you will gain practical experience in natural language processing and machine learning, learning how to preprocess text data, train a model, and assess its effectiveness.