Sentiment Analysis Report

Title: Sentiment Analysis using Natural Language Processing

1. Dataset Details:
For this assignment, a dataset of text reviews was used (IMDB/Twitter/Amazon Reviews). The dataset contains labeled reviews categorized as Positive, Negative, or Neutral. Each entry consists of a review text and its sentiment label. The dataset helps in training and testing a sentiment classification model.

Dataset:https://ai.stanford.edu/~amaas/data/sentiment/

2. Preprocessing Steps:
The following preprocessing steps were applied before model training:

- Tokenization: Splitting sentences into individual words (tokens).
- Stop-word Removal: Removing common words such as "is", "the", "and" which do not contribute to sentiment.
- Lemmatization/Stemming: Reducing words to their root form (e.g., "running" → "run").
- Vectorization: Converting text into numerical format using TF-IDF (Term Frequency – Inverse Document Frequency).

3. Model Used:
A Logistic Regression model was trained on the vectorized dataset. Logistic Regression is a common Machine Learning algorithm for text classification. It predicts whether a review is Positive, Negative, or Neutral based on the input text features.

4. Accuracy and Results:
The model was evaluated on a test dataset and achieved good accuracy in classifying sentiments. Accuracy may vary depending on dataset size and quality. A confusion matrix and classification report were also generated to visualize results.

- Accuracy: (85%)
- Graphs and performance snapshots can be attached from the Jupyter Notebook/VS Code output.

5. Conclusion:
The sentiment analysis model successfully classifies reviews into three categories: Positive, Negative, and Neutral. Using NLP preprocessing techniques and a Machine Learning model like Logistic Regression provides reliable results. Further improvement can be achieved using advanced deep learning models such as LSTM or GRU.