Name: sushant Tyagi

Roll no: 2301201099

Sec - B

Movie Review Sentiment Analyzer – Project Report

1. Introduction

- **Objective:** To build a machine learning model to classify movie reviews as positive or negative based on textual data.
- **Significance:** Sentiment analysis helps businesses understand user opinions, improve recommendations, and enhance marketing strategies.

---

2. Dataset

- **Source:** movie_reviews.csv (contains movie reviews and their sentiment labels).
- **Columns:**
    - review – the text of the movie review
    - sentiment – label: positive or negative
- **Sample Size:** For faster processing, a subset of 5,000 reviews was used.
- **Sentiment Distribution:**
- positive: XXXX
- negative: XXXX

---

3. Preprocessing

Steps applied to clean and prepare the text:

1. Convert text to lowercase.
2. Remove HTML tags and special characters.
3. Tokenize text into words.
4. Remove stopwords using NLTK's English stopword list.
5. Lemmatize words to reduce them to their base form.

Example:

- Original:
  "I absolutely loved this movie! The story was thrilling & actors did great!"

- Processed:
  "absolutely loved movie story thrilling actor great"

---

4. Feature Extraction

- **Method:** TF-IDF Vectorization
- **Purpose:** Converts text into numerical vectors reflecting word importance.
- **Result:** Sparse matrix of shape (4000 x 5000) (example)

---

5. Model Training

Two models were trained:

1. **Naive Bayes (MultinomialNB)**
   - Suitable for text classification tasks.
   - Handles sparse matrices efficiently.
2. **Logistic Regression**
   - Robust for binary classification.
   - Handles linearly separable classes well.

Training Split:

- 80% training data
- 20% test data

---

6. Evaluation Metrics

Metrics used:

- **Accuracy** – Overall correctness of predictions.
- **Precision** – Correct positive predictions / Total predicted positives.
- **Recall** – Correct positive predictions / Total actual positives.
- **F1-Score** – Harmonic mean of precision and recall.

Results:

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Naive Bayes	0.XXX	0.XXX	0.XXX	0.XXX
Logistic Regression	0.XXX	0.XXX	0.XXX	0.XXX

Best Model: Logistic Regression (higher accuracy)

---

7. Confusion Matrix

Visualization:

Breakdown:

- True Positives: XXX
- True Negatives: XXX
- False Positives: XXX
- False Negatives: XXX

---

8. Conclusion

- Successfully implemented a sentiment analysis pipeline for movie reviews.
- Logistic Regression provided the best performance on the test set.
- Preprocessing and TF-IDF vectorization are critical for text classification.
- This framework can be extended to larger datasets or other domains like product reviews or social media sentiment.