Project Overview

This project implements a **text classification pipeline** to predict the sentiment (Positive or Negative) of customer reviews using **Natural Language Processing (NLP)** techniques and a **Naive Bayes classifier**.
It includes preprocessing of raw text, feature extraction using **Bag of Words (CountVectorizer)**, model training, and performance evaluation.

Workflow Summary

1. Data Preprocessing
2. Text Cleaning and Normalization
3. Feature Extraction (Bag of Words)
4. Model Training (Multinomial Naive Bayes)
5. Performance Evaluation
6. Hyperparameter Tuning (Alpha)
7. Visualization (Confusion Matrix Heatmap)

Design Choice

- MultinomialNB is well-suited for **discrete word count features**.
- Fast, simple, and effective for text classification.
- nltk for natural language preprocessing (stopwords, stemming).
- sklearn for machine learning utilities and evaluation metrics.
- matplotlib & seaborn for visualization.

Challenges Faced

- Managing large vocabularies caused slow training — solved by limiting to 1500 features.
- Overfitting on small datasets — mitigated with train-test split.
- Finding optimal alpha value — required manual iteration.