Sentiment Analysis Project Report

1. Data Cleaning and Preprocessing:

- Removed special characters, numbers, and punctuation from review texts.

- Converted all text to lowercase.

- Tokenized text into individual words.

- Removed stop words (common but uninformative words like 'and', 'the').

- Applied stemming or lemmatization to reduce words to their base/root forms.

2. Sentiment Labeling (Optional):

- Initially used a pre-trained sentiment analyzer (like VADER) to label text.

- Later trained a custom classification model using labeled data.

3. Feature Extraction:

- Used TF-IDF (Term Frequency-Inverse Document Frequency) to convert text into numeric vectors.

- TF-IDF reflects how important a word is in a document relative to the entire dataset.

4. Model Training:

- Trained Logistic Regression and Naive Bayes classifiers.

- Used the TF-IDF features as input to the models.

- Split data into training and testing sets for evaluation.

5. Model Evaluation:

- Evaluated models using accuracy score and classification report.

- Best model (Logistic Regression) achieved approximately 91.5% accuracy.

Classification Report:

	precision	recall	f1-score	support
Negative	0.88	0.98	0.93	95
Neutral	0.00	0.00	0.00	4
Positive	0.96	0.89	0.92	101
Accuracy			0.92	200
Macro Avg	0.61	0.62	0.62	200
Weighted Avg	0.90	0.92	0.91	200

==========