**Sentiment Analysis on Reviews**

This project focuses on predicting sentiment based on text reviews using machine learning. A complete preprocessing and modeling pipeline is implemented to achieve high accuracy.

**Dataset**

* The dataset is sourced from **Kaggle**.
* It contains text reviews along with their corresponding sentiment labels (positive/negative).

**Project Pipeline**

1. **Upload Dataset**  
   The dataset is uploaded and loaded into the environment for analysis.
2. **Text Preprocessing**
   * **Convert to lowercase**: All text in the review column is converted to lowercase for consistency.
   * **Remove punctuation**: All punctuation characters are removed.
   * **Remove numbers and emojis**: Any digits and emojis are removed from the text.
   * **Remove stopwords**: Common stopwords like “the”, “is”, etc., are removed.
     + Two methods were considered: **tokenization** and **split method**.
     + The split method had limitations since we needed to ensure punctuation was removed before splitting.
3. **Train-Test Split**  
   The dataset is split into training and testing sets for model evaluation.
4. **Feature Extraction**  
   Text data is converted into numerical vectors for machine learning using:
   * **CountVectorizer**
   * **TF-IDF Vectorizer**
5. **Model Training**  
   Two models are trained on the extracted features:
   * **Multinomial Naive Bayes (MultinomialNB)**
   * **Logistic Regression**

**Results**

| **Feature Extraction** | **Model** | **Accuracy** |
| --- | --- | --- |
| CountVectorizer | MultinomialNB | 85% |
| TF-IDF Vectorizer | MultinomialNB | 86% |
| TF-IDF Vectorizer | Logistic Regression | 89% |

**Conclusion**

* **Best performance** is achieved using **TF-IDF Vectorizer with Logistic Regression**.
* Proper text preprocessing and feature extraction significantly improve model performance.