Sentiment Analysis of Real-time Flipkart Product Reviews

Project Overview :

Online product reviews play a critical role in influencing customer decisions and business strategies. This project focuses on analyzing customer reviews from Flipkart to determine whether a review expresses positive or negative sentiment.

The goal is to build an end-to-end sentiment analysis system, starting from data preprocessing to model deployment on the cloud using AWS EC2.

Objectives :

- Analyze customer review text to understand sentiment
- Clean and preprocess textual data
- Convert text into numerical representations
- Train and evaluate machine learning models
- Deploy the trained model as a web application
- Enable real-time sentiment prediction for user-entered reviews

Dataset Description :

The dataset consists of Flipkart product reviews containing:

- Review text
- Sentiment labels (positive / negative)

The data was used to train and evaluate sentiment classification models.

Workflow and Implementation :

1. Data Loading and Exploratory Data Analysis

- Loaded the dataset into a Jupyter Notebook
- Checked dataset shape, missing values, and class distribution
- Analyzed the balance between positive and negative reviews

2. Data Cleaning and Preprocessing

The following preprocessing steps were applied:

- Converted text to lowercase
- Removed punctuation and special characters
- Removed stopwords
- Tokenized the text
- Normalized review content

This step ensured clean and meaningful input for model training.

3. Text Embedding (Feature Engineering)

- Used TF-IDF (Term Frequency–Inverse Document Frequency) to convert text into numerical vectors
- TF-IDF was chosen due to its effectiveness, simplicity, and strong performance for text classification tasks

4. Model Training

Multiple machine learning models were trained and compared:

- Logistic Regression (final selected model)

The dataset was split into training and testing sets before model training.

5. Model Evaluation

- The models were evaluated using the F1-score
- Logistic Regression achieved the best balance between precision and recall
- The final model was selected based on performance metrics

6. Web Application Development

- Developed a Streamlit web application
- Users can enter a product review and receive instant sentiment prediction
- The application loads:

1. Trained model (sentiment_model.pkl)
2. TF-IDF vectorizer (tfidf_vectorizer.pkl)

7. Model Deployment on AWS EC2

- Created an AWS EC2 Ubuntu instance
- Installed required dependencies (Python, pip, virtual environment, Streamlit)
- Uploaded model files and Streamlit app code
- Configured security groups to allow access on port 8501
- Deployed and ran the Streamlit application on the EC2 instance

8. Testing and Monitoring

- Tested the application using multiple positive and negative reviews
- Verified real-time predictions through the web interface
- Ensured the application works correctly after deployment

Technology Stack :

- **Programming Language:** Python
- **Libraries:**
  - pandas, numpy
  - scikit-learn
  - nltk
  - streamlit
- **Machine Learning Model:** Logistic Regression

- **Text Representation:** TF-IDF
- **Cloud Platform:** AWS EC2
- **Development Tools:** Jupyter Notebook, VS Code

Results

- Successfully built an end-to-end sentiment analysis system
- Achieved reliable sentiment predictions using TF-IDF and Logistic Regression
- Deployed a real-time, cloud-hosted web application accessible via browser

Conclusion

This project demonstrates a complete machine learning pipeline, from data preprocessing and model training to deployment on a cloud platform. The system can help businesses analyze customer feedback efficiently and make data-driven decisions.

The deployed Streamlit application provides an easy-to-use interface for real-time sentiment analysis.