Twitter Sentiment Analysis – Explanation

Overview

This project performs Sentiment Analysis on Twitter data using the Sentiment140 dataset. The goal is to classify tweets as either Positive or Negative based on their content. The steps include data acquisition, preprocessing, feature extraction, model training, and evaluation.

Setup and Dataset Download

The project begins by installing and configuring the Kaggle API to download the Sentiment140 dataset. It contains 1.6 million labeled tweets. After downloading, the zip file is extracted for use.

Importing Libraries

Libraries like pandas, numpy, nltk, and scikit-learn are imported for data processing, text handling, and machine learning. Stopwords are also downloaded for text cleaning.

Data Loading and Inspection

The dataset is loaded with pandas and assigned column names such as 'target', 'id', 'date', 'flag', 'user', and 'text'. Basic inspection is done to check shape, missing values, and sentiment distribution.

Data Preprocessing

Data is cleaned and preprocessed. Sentiment values are converted to binary (0=negative, 1=positive). Text is cleaned by removing non-alphabetic characters, converting to lowercase, removing stopwords, and applying stemming using NLTK's PorterStemmer.

Feature Extraction

Cleaned tweets are transformed into numerical features using TF-IDF Vectorizer. This technique measures the importance of words in relation to the dataset, producing numerical vectors for model training.

Model Training

The dataset is split into training and testing sets. Logistic Regression is used as the classifier because it's effective for binary sentiment tasks. The model learns relationships between word features and sentiment labels.

Model Evaluation

The model’s performance is tested using accuracy score. The accuracy on training and testing data indicates how well the model generalizes to unseen tweets.

Prediction on New Data

The trained model can predict sentiment for new tweets. Example: ‘I love this product!’ results in a positive label, while ‘This is terrible’ yields a negative label.

Conclusion

The project demonstrates the complete sentiment analysis workflow — from raw data to prediction. Using Logistic Regression with TF-IDF provides a strong baseline.