SENTIMENT ANALYSIS

Project: Sentiment Analysis of Tweets

Overview

I developed a machine learning model to classify the sentiment of tweets into positive, negative, or neutral categories. This project involved various stages, including data cleaning, feature engineering, exploratory data analysis, model selection, evaluation, and deployment using Python and key data science libraries. The goal was to build an accurate and efficient predictive model to provide insights into the factors influencing sentiment.

Problem Statement

The objective of this project was to classify the sentiment expressed in tweets using natural language processing and machine learning techniques. Given the tweet text and additional metadata, we aimed to predict whether the sentiment was positive, negative, or neutral.

Tools & Technologies

- **Python:** The primary programming language used.
- **Pandas & NumPy:** For data manipulation and analysis.
- **NLTK & SpaCy:** For natural language processing tasks.
- **Scikit-Learn:** For building and evaluating machine learning models.
- **Matplotlib & Seaborn:** For data visualization and exploratory data analysis.
- **Flask:** For deploying the model as an interactive web application.

Data Preparation

- **Data Cleaning:** Handled missing values and inconsistencies in the dataset. For example, filled missing text values with empty strings.
- **Text Preprocessing:**
  - **Cleaning:** Removed HTML tags, non-alphabetic characters, and converted text to lowercase.

  - **Tokenization and Stopword Removal:** Tokenized the text and removed stopwords using NLTK.

- **Feature Engineering:** Created new features such as:

  - **TF-IDF Features:** Converted cleaned and preprocessed text into numerical feature vectors using TF-IDF.

- **Encoding Categorical Variables:** Converted categorical variables (e.g., sentiment) into numerical format using label encoding.

Exploratory Data Analysis (EDA)

- Visualized distributions and relationships between features using histograms, box plots, and heatmaps.

- Identified key factors affecting sentiment, such as the occurrence of certain words and phrases.

Model Development

- **Algorithm Selection:** Evaluated multiple machine learning algorithms including:

  - Logistic Regression

  - Naive Bayes

  - Random Forest

  - Support Vector Machines (SVM)

- **Model Tuning:** Used cross-validation and grid search to optimize hyperparameters and improve model performance.

- **Evaluation Metrics:** Assessed models using accuracy, precision, recall, and F1-score to ensure balanced and robust performance.

Results

- **Performance:** The final model, a Random Forest classifier, achieved:

  - **Accuracy:** 78.21%

- **Key Insights:**

  - **Words and Phrases:** Certain words and phrases were strong indicators of positive or negative sentiment.

  - **Text Length:** Longer tweets tended to have more neutral sentiment.

  - **Time of Tweet:** Time of day had a minor influence on the sentiment distribution.

Key Learnings

- **Data Preprocessing:** The importance of handling missing data, text cleaning, and creating meaningful features to improve model accuracy.

- **Model Evaluation:** Understanding the trade-offs between different performance metrics and selecting the best model based on balanced performance.

- **Deployment:** Gaining practical experience in deploying a machine learning model and creating an interactive user interface for real-world applications.

GitHub Repository

Explore the complete project and code here:

https://github.com/sravanthi224/Quanta-3.git