🐦 Twitter Sentiment Analysis

Project Overview

The **Twitter Sentiment Analysis** project aims to classify tweets into **positive**, **negative**, or **neutral** sentiments. Understanding public sentiment on social media helps **businesses**, **policymakers**, and **researchers** gauge opinions, monitor trends, and make informed decisions.
This project utilises **Natural Language Processing (NLP)** and **Machine Learning** techniques to perform **sentiment classification** on real Twitter data.

Dataset

The dataset consists of real-world tweets labelled with sentiment categories. It includes user-level and tweet-level metadata.

Key Columns:

- `twitter_id` – Unique identifier for each tweet
- `airline_sentiment` – Target variable (positive/negative/neutral)
- `text` – The tweet text content
- `retweet_count` – Number of times the tweet was retweeted
- `airline` – Airline company associated with the tweet
- Other metadata: `tweet_coord`, `tweet_created`, `tweet_location`, `user_timezone`

Data Preprocessing

To prepare the textual data for analysis, the following **NLP preprocessing** steps were applied:

1. **Text Cleaning** – Removed URLs, punctuation, numbers, and special characters.
2. **Lowercasing** – Converted all text to lowercase to ensure uniformity.

3. **Tokenisation** – Split sentences into individual words (tokens).
4. **Stopword Removal** – Removed common words like *“the”, “is”, “that”* that do not add meaning.
5. **Vectorization** – Transformed text into numerical features using **TF-IDF** or **Bag of Words**.
6. **Train-Test Split** – Divided the dataset into training and testing sets for model evaluation.

Machine Learning Models

Several classification models were trained and compared to determine the best performer:

- **Logistic Regression** – A simple and efficient linear model for multi-class classification.
- **Multinomial Naive Bayes** – Probabilistic model well-suited for text classification.
- **Support Vector Classifier (SVC)** – Handles high-dimensional text features effectively.
- **Random Forest Classifier** – Ensemble model combining multiple decision trees for higher accuracy and robustness.

Model Evaluation Metrics

Performance was evaluated using standard classification metrics:

- **Accuracy** – Overall correctness of predictions.
- **Precision** – Proportion of correctly predicted sentiments among all predictions of that sentiment.
- **Recall** – Proportion of actual sentiments correctly identified by the model.
- **F1-Score** – Balance between precision and recall.
- **Confusion Matrix** – Visualisation of correct and incorrect predictions across sentiment classes.

Insights

Key findings and observations derived from the analysis:

- **Negative tweets** mostly discussed **service issues and delays**.
- **Positive tweets** reflected **satisfaction and appreciation** for good service.
- **Neutral tweets** were mostly **informational or general statements**.
- **Logistic Regression** and **Random Forest** provided the best overall performance and interpretability.

Skills Learned

Throughout this project, the following technical and analytical skills were developed:

- Text preprocessing and data cleaning using NLP techniques
- Feature extraction from text using **TF-IDF** and **Bag of Words**
- Applying **multi-class classification** algorithms
- Evaluating model performance using **precision, recall, F1-score, and confusion matrix**
- Drawing actionable insights from **social media sentiment data**

Technologies Used

- **Python** (pandas, numpy, scikit-learn, nltk, re, matplotlib, seaborn)
- **Natural Language Processing (NLP)** techniques
- **Jupyter Notebook / Google Colab**
- **Machine Learning Models** (Logistic Regression, Naive Bayes, SVC, Random Forest)