Project Report: Airline Sentiment Analysis Using NLP & Machine Learning

---

Project Title:

Airline Sentiment Analysis Using NLP & Machine Learning

Business Problem

Airlines often receive real-time feedback from customers via social media platforms like Twitter. However, with the volume of tweets, it becomes challenging to manually analyze them for sentiment and actionable insights.

Objective: To automatically classify tweets about airlines into sentiment categories (positive, neutral, negative) using Natural Language Processing (NLP) and Machine Learning (ML). Additionally, extract insights on common customer complaints and sentiment trends.

Dataset Overview

- **Source:** Twitter US Airline Sentiment Dataset (Kaggle)
- **Rows:** ~14,500
- **Columns Used:** `airline`, `airline_sentiment`, `text`, `negativereason`

Column Name	Description
airline	Name of the airline
airline_sentiment	Sentiment label (positive, neutral, negative)
text	The tweet text
negativereason	Reason for negative sentiment (if applicable)

Methodology

1. Data Cleaning & Preprocessing

- Removed URLs, hashtags, mentions
- Removed punctuation and numbers
- Lowercased all text
- Stopwords removal using NLTK
- Lemmatization using WordNet

2. Feature Engineering

- Created a new column `clean_text` for modeling
- Encoded sentiments as integers (0: Negative, 1: Neutral, 2: Positive)

3. Text Vectorization

- Used **TF-IDF Vectorizer** with n-gram range (1,2)
- Limited to top 5000 features to balance performance and accuracy

4. Model Selection

- Chose **Multinomial Naive Bayes** due to its effectiveness for text classification

5. Model Training

- 80% Train / 20% Test split
- Trained on TF-IDF features

6. Model Evaluation

- **Accuracy:** 74.04%
- **Classification Report:**
  - Negative: Precision 0.73, Recall 0.97
  - Neutral: Precision 0.70, Recall 0.29
  - Positive: Precision 0.86, Recall 0.42
- **Confusion Matrix:**
  - High precision for Positive sentiment
  - Neutral sentiment classification needs improvement

Visualizations

Page 1: Model Metrics

- Confusion Matrix
- Classification Report

Page 2: Sentiment Distribution by Airline

- Bar chart grouped by airline and sentiment

Page 3: Negative Sentiment Analysis

- Horizontal bar chart of top 10 `negativereason`

Page 4: Sentiment Trends Over Time

- Simulated time-based line plot (resampled hourly tweets to daily aggregates)

Page 5: Streamlit App

- Web app for real-time tweet sentiment classification

Tools & Technologies

- **Language:** Python
- **Libraries:** pandas, numpy, scikit-learn, nltk, seaborn, matplotlib
- **Deployment:** Streamlit
- **Version Control:** Git & GitHub

Business Impact

- Identify top pain points from customer feedback
- Improve airline service quality and customer engagement
- Benchmark competitors via social media sentiment

- Real-time alert systems during high-volume negative feedback

---

Future Enhancements

- Integrate with Twitter API for live data ingestion
- Use deep learning models (LSTM, BERT)
- Develop a Power BI dashboard for executive summary
- Incorporate multilingual sentiment analysis

Author

Eyesly Meribha Johnson Paulraj
MSc Data Science | Data Scientist
Email: eyesly.meribha.jp@gmail.com
