Project Proposal: Twitter Airline Sentiment Analysis

1. Dataset:

Data Source: Twitter Airline Sentiment Dataset https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment

Description: This dataset contains tweets related to airline experiences, categorized into three sentiment classes: positive, negative, and neutral. It includes information such as tweet text, sentiment label, and airline information. The tweets in this dataset are from 2015, which may impact the relevance of the sentiment analysis given the potential changes in airline industry sentiment over time.

Additionally, we recognize the importance of fresh data, and we shall make an effort to obtain more recent tweets related to airline experiences to supplement our analysis.

2. Project Objectives:

The primary objective of this project is to perform sentiment analysis on tweets related to airline experiences to gain insights into customer sentiments. Due to the age of the dataset (tweets from 2015), we acknowledge the limitation and intend to address the following research questions:

Sentiment Distribution: What is the distribution of sentiment (positive, negative, neutral) in tweets about airline experiences?

Airline Performance: How do different airlines fare in terms of sentiment? Are there airlines with consistently better or worse sentiment?

Temporal Trends: Are there any temporal trends in sentiment, such as seasonality or changes over time?

Key Topics: What are the most common topics or issues mentioned in negative tweets about airlines?

Sentiment Impact: How does sentiment correlate with factors like retweets, likes, and reply counts?

Geographic Variation: Is there geographic variation in sentiment? Do certain regions exhibit more positive or negative sentiment toward airlines?

3. Methodology:

Data Preprocessing:

Data cleaning and text preprocessing to remove noise, handle missing values, and tokenize tweet text.

Sentiment Analysis:

Utilize Natural Language Processing (NLP) techniques to perform sentiment analysis on tweet text. Common libraries like NLTK or spaCy will be used for this task.

Exploratory Data Analysis (EDA):

Conduct EDA to explore the sentiment distribution, temporal trends, and geographic variations.

Airlines Comparison:

Compare sentiment scores for different airlines to identify top-performing and poorly-performing airlines.

Topic Analysis:

Use techniques such as topic modeling to identify common topics in negative tweets.

Correlation Analysis:

Analyze correlations between sentiment scores and engagement metrics like retweets, likes, and reply counts.

Geospatial Analysis:

Employ geospatial visualization tools to visualize geographic variations in sentiment.

4. Expected Outputs:

A detailed report summarizing the findings and insights from the analysis, including sentiment distribution, airline performance, temporal trends, key topics in negative tweets, and more.

Visualizations, including sentiment distribution charts, temporal trend graphs, airline comparison plots, and geospatial heatmaps.

A presentation summarizing the key findings for easy communication.

Python code and Jupyter notebooks for reproducibility.

This project proposal outlines a comprehensive plan to analyze sentiment in airline-related tweets, providing valuable insights into customer opinions and experiences with airlines. We acknowledge the age of the dataset and the importance of obtaining fresh data to enhance our analysis. The methodology includes various data analysis techniques to answer the specified research questions.