PROJECT DOCUMENTATION

SENTIMENT ANALYSIS FOR MARKETING

# **INTRODUCTION :**

Analyzing the sentiment of airline tweets can provide valuable insights into customer satisfaction, brand perception, and emerging trends. By understanding what customers are saying about airlines on social media, businesses can improve their marketing campaigns, develop new products and services, and better understand customer needs.

# **ABOUT DATASET :**

The dataset used in the code you provided is a CSV file called "Tweets.csv". The dataset contains the following columns:

* tweet\_id: The unique identifier for the tweet.
* airline\_sentiment: The sentiment of the tweet towards the airline, which can be either "positive", "negative", or "neutral".
* airline\_sentiment\_confidence: The confidence score of the airline sentiment prediction.
* negative\_reason: The specific reason for the negative sentiment, if any.
* negative\_reason\_confidence: The confidence score of the negative reason prediction.
* airline: The name of the airline.
* airline\_sentiment\_gold: The gold standard airline sentiment, which is used to evaluate the performance of the airline sentiment prediction model.
* negative\_reason\_gold: The gold standard negative reason, which is used to evaluate the performance of the negative reason prediction model.
* retweet\_count: The number of times the tweet has been retweeted.
* text: The text of the tweet.
* tweet\_coord: The coordinates of the tweet, if available.
* tweet\_created: The date and time at which the tweet was created.
* tweet\_location: The location of the tweet, if available.
* user\_timezone: The timezone of the user who created the tweet.

# **PREPROCESSING STEPS:**

* Remove punctuation and special characters: This will make the text data more consistent and easier for the machine learning model to process.
* Tokenize the text: This will split the text into individual words or tokens.
* Lemmatize or stem the text: This will reduce the words to their root form, which can help the machine learning model to identify similar words.
* Convert the categorical data to numerical data: This can be done using one-hot encoding or label encoding.
* Split the data into training and test sets: This is done to prevent overfitting and to get an unbiased evaluation of the model's performance.

**Choice of ML algorithm:**

A variety of machine learning algorithms can be used for sentiment analysis and negative reason prediction. Some common choices include:

* Naive Bayes: Naive Bayes is a simple but effective algorithm that is well-suited for text classification tasks.
* Logistic regression: Logistic regression is another simple but effective algorithm that can be used for binary classification tasks, such as predicting whether a tweet is positive or negative.
* Support vector machines (SVMs): SVMs are a powerful machine learning algorithm that can be used for both classification and regression tasks.
* Neural networks: Neural networks are a type of deep learning algorithm that can be used for a variety of tasks, including sentiment analysis and negative reason prediction.

**Model training:**

Once the data has been preprocessed, the machine learning model can be trained. This is done by feeding the training data to the model and allowing it to learn the relationships between the features and the target variable.

Evaluation metrics:

The performance of the machine learning model can be evaluated using a variety of metrics, including:

* Accuracy: Accuracy is the percentage of predictions that the model makes correctly.
* Precision: Precision is the percentage of positive predictions that are actually positive.
* Recall: Recall is the percentage of actual positive cases that are correctly predicted by the model.
* F1 score: The F1 score is a harmonic mean of precision and recall.

**Innovative technique used:**

One innovative technique that can be used for sentiment analysis and negative reason prediction is to use a deep learning model with a bidirectional encoder representation from transformers (BERT) architecture. BERT is a type of language model that has been shown to achieve state-of-the-art results on a variety of natural language processing tasks.

**RESULT:**

This code performs sentiment analysis and negative reason prediction using a BERT model on a dataset of tweets. The data is preprocessed by removing punctuation and special characters, tokenizing the text, lemmatizing the text, and converting the categorical data to numerical data. The model is trained on the encoded data and evaluated using accuracy, precision, recall, and F1 score.