**Sentiment Analysis**

**Data Pre-processing:**

1. Removed **@’Flight\_Name’** from the beginning of the text and stored ‘Flight\_Name’ in a new column.

2. Stored positive or negative **emotion values** in a new column corresponding to the emoticons.

3. Removed **stopwords** and **punctuations** and ignored non-letter text (numerical values, hyperlinks).

4. Normalised text into **tokens** after cleaning and applying **lemmatisation** for better results.

5. Applied **vectorisation** on the normalised text to map text to vectors for model applicability.

6. Mapped sentiment values to **numerical values** (Positive -> 0 and Negative -> 1).

**Approaches:**

**1. Machine Learning:**

Applied well-known classification algorithms such as Support Vector Machine (SVM), Naïve Bayes (Multinomial & Gaussian), AdaBoost, Random Forest and Bagging classifier with a base estimator. Improved scoring metrics using Grid Search on parameter space of the particular classifier.

**SVM (C = 10, ‘gamma’ = ‘scale’, kernel = ‘rbf’)** gave the best results:

**Train:** Precision – 0.9979, Recall – 0.9995, Accuracy – 0.9995, F1 Score – 0.9987

**Test:** Precision – 0.6679, Recall – 0.8860, Accuracy – 0.9207, F1 Score – 0.7681

**2. Deep Learning:**

Trained a **Bidirectional LSTM** having an initial layer of word embedding. That is followed by a 1D Convolution layer and Max Pooling to reduce the feature map. A dropout layer is also added to avoid any overfitting. The activation function chosen is ‘sigmoid’ because of binary classification, and the loss function is ‘binary\_crossentropy’.

**Train:** Precision – 0.9901, Recall – 0.9928, Accuracy – 0.9863, F1 Score – 0. 9915

**Test:** Precision – 0.9288, Recall – 0.9662, Accuracy – 0.9129, F1 Score – 0.9440

**3. Transformers:**

Tokenised the text using the pre-trained **‘bert-base-uncased’** tokeniser. Performed encoding using the ‘encode\_plus’, which returns a dictionary containing the encoded sequence and mask for sequence classification.

**Test:** Precision – 0.8930, Recall – 1, Accuracy – 0.9775, F1 Score – 0.9435

**Comments:**

Since the dataset is **imbalanced** (Positive sentiments = 2363, Negative sentiments = 9178), accuracy is not a suitable scoring parameter instead, the **F1 Score** should be considered. **Data augmentation** can be done as a future course of action (Back Translation, NLP Albumentation, NLPAug Library) to restore data balance.