# Sentiment Analysis Prediction Project:

**Problem Statement**:

Build a sentiment analysis model to analyze customer reviews and feedback for a hospitality company. The model should be able to classify reviews as positive, negative.

**Dataset**: <https://www.kaggle.com/code/humzafazalabbasi/sentiment-analysis-hotel-reviews/notebook>

Dataset was taken from Kaggle on Hotel reviews, It consist of Two feature Review(text) and Rating(numerical) and 20491 rows.

1. **Separated** data input variable is Review and Target variable is Rating.

2. **Splitting** the data into train and test as 75:25 ratio

3. Data Preprocessing: To clean the unnecessary data we have using some Techniques

(i) Regex: to the remove unimportant information the Review data like(numerical, Special char.

(ii)Tokenize: for split the text into word, Lower(): to convert into Lower case, stop words removal

(iii)Stemming and Lemmatizing to reduce the Vocab size and which is used to reduce the dimension

4. **Data Transforms**: as Problem statement used to build the machine Model, I used the BOW(count Vector), TD-IDF (Frequency based Vector) .

Fit and Transform for the Train data and only transformer for the test data.

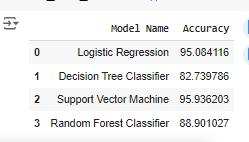
5**. Model Building**: As Problem statement and Dataset, it comes under the Classification Task,

So, I have used the Algorithms like, Logistics Regression which will performer for Binary classifier,

Ensemble Algorithms are Support Vector Machine Classier, and have used one more Algorithm is Decision Tree.

6. **Prediction and evaluation**: Prediction the model by using the test data, and Evaluating the Model by using the Actual and Predict as metrics Like accuracy, Confusion matrix.

7. **Conclusion**: Can be observed that, the accuracies of the I similar using both the Preprocessing techniques, The Logistic Regression and Support Vector Machine has the highest accuracy of 95% compared to the other models.



Issue: 1. Data was imbalance from which we can used the oversample and under sample, if we oversample and under sample, the will be in more, we can use POS