**Problem Statement** – Analyze the review comments of the customers of a hotel and predict if the ratings given are ‘Good’ or ‘Bad’.

**Tools Used** – R programming language is used mainly due to the large number of text mining packages available in R and, also due to the ease of data cleaning features offered.

**Approach/Steps Followed** –

1. Hypothesis Generation
2. Exploratory Data Analysis (EDA)
3. Data Preparation (Cleaning, Oversampling to avoid data imbalance, Feature Engineering)
4. Modeling
5. Prediction and Parameter tuning
6. Result File Creation
7. Hypothesis Generation

Before starting actual coding, hypothesis was generated based on the available domain knowledge and, also based on the information collected from internet. It was evident that problem would have text mining and natural language processing as we need to deal with review comments. Few key ideas/entities that came in during hypothesis generation phase are,

1. User Id
2. Review Date
3. Review Rating
4. Review Title
5. User City
6. Will Recommend or Not
7. Reviews
8. Device used for uploading review
9. EDA

After loading data in R environment, next step was to do a quick data exploration to understand the data. Univariate and Multivariate analysis is done. On exploring,

1. It is identified that we have just 3 key features (Browser\_Used, Device\_Used, Description) and the response variable (Is\_Response). It is evident from the features that Description plays an important role in prediction.
2. Data cleaning is needed in ‘Browser\_Used’ column as same browser names are represented in different ways.
3. Data seems to have class imbalancement with 68.17% is Good and 31.83 is Bad.
4. Data Preparation (Cleaning + Feature Engineering)
5. ‘Browser Used’ column was merged to form unique levels
6. Data is over sampled to bring ‘Bad’ class to match with the number of ‘Good’ class.
7. New columns are created to represent Description column length and Description word count.
8. Couldn’t try detailed data cleaning because of the time constraint I fall in due to some unplanned personal emergency.
9. Modeling
10. Corpus for description column was created and, preliminary cleansing is done to get data ready for mining and prediction.
11. Couldn’t try multiple algorithm because of the time constraint I fall in due to some unplanned personal emergency.
12. Xgboost algorithm is used to fit the model
13. Prediction & Parameter tuning
14. Predicted test data based on the fitted model.
15. Couldn’t do hyper-parameter tuning because of the time constraint I fall in due to some unplanned personal emergency.
16. Result file creation
17. Saved the results data into **xgb\_results.csv** file.