**Case Study 4**

'Suppose you are working as compliance data science analyst in a firm. One of your task is to monitor internal communications in order to better understand employees’ moods and assess any potential risks. You can leverage sentiment analysis technology which has become a form of risk management and is emerging as a useful risk control tool for a variety of businesses to identify and address regulatory risk issues, compliance problems and potential fraud.'

**Dataset Description:**

You have received a dataset which contains 3 columns:

ID (Unique identifier for a sentence)

label (1-Positive & 0-Negative)

Text (Sentences)

**Evaluation Criteria:**

You can leverage R or Python programming language for solving the assignment.

**Programming best practices:**

You need to explicitly import all the packages/libraries at the starting of the notebook.

For any non-trivial pieces of code, please write supporting comments.

All the cells should run with no error.

**Submission guidelines:**

You need to zip and attach:

Final complied script.

Using the data Practise Test Case 4 Data.xlsx, solve the following questions:

1. Read the dataset and print the schema
2. How many Unique Ids are there in the dataset? (\*Look for exact answer)
3. How many total missing values are there across all columns in the dataset? (\*Look for exact answer)
4. How many rows are there in the dataset after dropping the rows containing the missing values and removing duplicates (keep first occurrence in case of duplicates)? (\*Look for exact answer)
5. What is the percentage of positive labels in the dataset? (\*Look for exact answer)
6. How many unique tokens are present in the dataset including stop words? (\*Look for code and analysis)
7. How many unique tokens are present in the dataset excluding English stop words? (\*Look for code and analysis)
8. What is Zipf distribution and Plot Zipf distribution for 500 tokens in decreasing order including stop words? (\*Look for descriptive answer, code and plot)
9. Plot frequency distribution of top 50 tokens in negative data including stop words
10. Plot frequency distribution of top 50 tokens in positive data including stop words
11. If you have to predict the sentiments then explain the ML life cycle in detail?
12. Identify the suitable embedding for input features and build a Machine Learning model to predict the sentiments?
13. Identify the suitable Model validation matrix and explain why you have chosen the same?
14. What are the model serialization techniques?
15. Leverage the best serialization technique and save your model?