**INSTRUCTIONS TO WORK ON DATASETS**

**Dataset Description:**

* Train & Val Dataset contains 1000 rows and 2 columns i.e. Conversations & Patient Category (0- Not a patient , 1- Patient)
* Patient Category is the target variable
* Test Dataset contains 157 rows and 1 column (Conversations)

**Instructions:**

* Agenda is prediction of the Target Variable (Patient Category) in Test Dataset
* Candidates are not allowed to skip any of the steps mentioned below; additionally, candidates can incorporate any new steps / methods to carry out the task.
* Also, candidates have full freedom to use Python Libraries or Algorithms or Visualizations of their choice to achieve final goal.

**Steps to carry out**

**1. Text Processing:**

Remove Unwanted Digits

Use Lowercase

Remove Diacritics

Remove Stopwords

Remove Punctuation

FillNAs

Removes white space

Remove URLs

Remove HTML tags

Remove Contractions

Simple normalization — (e.g. standardize near identical words)

Advanced normalization (e.g. addressing out-of-vocabulary words)

Spell or grammar check

Noise % Calculation before & after text processing

Stemming

POS & Lemmatization

Tokenization

**2. Extract Information:**

Extract Email & NER

**3. Sentiment Analysis:**

Sentiment Analysis & its interpretation

**4. Tables:**

Word Frequency Table

Word Frequency – Normalization with the maximum frequency available

Sentence Frequency Table

Sentence Frequency - Normalization with the maximum frequency available

**5. Features:**

n-grams (1-4);

Map words into Vectors, TF-IDF,word to vec, Glove(any of these)

Kmeans to cluster based on the content and tfIdf, with support words in table or visuals

Perform PCA to increase interpretability & reduce info loss

**6. Visualizations**:

Visualize Top 20 words & Bottom 20 Words (representative of the class)

Wordcloud

Sentiment polarities

Visualize Kmeans & PCA (tfidf) results

TNSE plot of the model

**7. Summarization:**

Extractive Summary - 30% of the total content for each row

Abstract Summary - 30% of the total content for each row

General Summary Comparison between 3 libraries at the least

**8. Modeling & Tuning:**

Train & Val Dataset [Train-80%, Validation-20%]

Test Dataset [prediction of target variable]

ML Pipeline

Hyper Parameter tuning () & result interpretation

F1-score for each class

Micro Averaged F1 scores

Other Metrics & confusion matrix

Prediction using test data

Prediction Comparison

**9. Model Deployment:**

UI/user-friendly method for end user to access the predictive model