Step by step detailed write up of the entire approach-

1) As the first step we are importing the basic libraries which we will go adding on with our exploratory data analysis

2) Then we import the train and test datasets. Since they are not encoded in UTF-8,we specified the encoding explicitly

3) Then we explore the datasets by checking out their dimensions using .shape, .columns and head()

4) On exploring the columns in test dataset, we find it is having two extra columns 'Index' and 'Unnamed: 9',which were not present in the training dataset. They were nowhere relevant for our analysis part, hence we drop both of the columns.

5) Exploring the datasets and extracting more information from it,we find that most of the data is non-numeric in nature.

There were some missing values in the train and test set.The specific no. being as below-

Host-59

time(GMT)-161

Title-216

TRANS\_CONV\_TEXT=1

6) From the exploration we find that ,the maximum information of the dataset lies in the TRANS\_CONV\_TEXT column values which will help us to determine the labels of the text correctly.

The baseline model can be built just by using two columns - TRANS\_CONV\_TEXT and Patient\_Tag

Conscisely putting the problem statement of this binary classification task-

Given any conversation text, we need to determine whether the text is patient tagged or not.

7) For the single missing value in TRANS\_CONV\_TEXT, we apply imputation by mode to fill the value in.

8) In the next step, extracting the values of TRANS\_CONV\_TEXT and Patient\_Tag and separating them out in a Dataframe.

Saving the new dataframes for further use.

9)After checking the class distribution using pd.value\_counts upon the Patient\_Tag ,we found about the class imbalance.

But we do ignore it and try making the baseline model.

PS: In case of handling the class\_imbalance part we can use SMOTE analysis(Synthetic Minority Over Sampling Technique),but we have avoided using it here to carry out our analysis on the given dataset.

10) We start with the data preprocessing part where we do the following preprocessing steps-

-Converting to lower case

-Lemmatization

-Removal of digits

11) In this step after preprocessing we opt for Count-Vectorizing the text

12) Now for building the baseline and evaluation,we try it out with different models like-

Logistic Regression,XGBclassifier,Bernoulli NB and Random forest classifier.

13) The below are the accuracy values for the four models respectively-

Logistic Regression-91.3%

XGBclassifier-89.3%

Bernoulli NB-89.6%

Random forest classifier-85.9%

14) In the last step, we have to make the predictions and prepare the submission file

15) We proceed with repeating the same preprocessing steps with the test dataset

16) The submission , initially in the form of a data set is read into a csv file.