1. Problem :

Step1 : Found out label distribution - to check is it imbalance or skewed on single label
{
'booking': 4470,
'cancelation': 222,
'issues': 23,
'negotiation': 793,
'other': 1654,
'rebooking': 568
}

Used **F-Measure** as my performance measure

Step 2: Checked if any features were missing or invalid entries

Step 3: Analysed the samples and created the feature vector (numerical features)
Following are the features have used :

Feature for each token

All are binary features - Did not use length feature.
#where_1 = is_body
#where_2 = is_subject
#shape_1 = begins_with_capital
#shape_2 = contains_colon
#shape_3 = contains_hyphen
#shape_4 = contains_d (for date)
#start = is_begining (is 1 if its positon is less than 10)
#ner = 2 placeholder for every nerType (so 2*24) + 1 for other
#feature vector for every token = 2+4+1+49 = 56

Step 4: Since samples are of variable number of tokens, i padded all the samples to same number of tokens. (2600 token size)

Step 5: Divided my training samples into 2 sets- training and validation set

Step 6: Created the sparse feature matrix

Step 7: Trained using SVM on Training data (one vs all classifier startegy)

Step 8: Got the predicted probabilities on Validation dataset

Step 9: Tuned the threshold for every class independently (to increase the performance measure)

Step 10 : For every sample, got the probability values for each class and if the probability value is greater than the given threshold, then label is assigned to that sample

Q2. How do you evaluate your performance and how well do you perform?

Used **F-measure** as performance measure.
On validation dataset: (25 percent of random data)

Note : I used random sampling, **could have used stratified sampling**

F-score on each class on validation dataset is as follows:

Other	0.631578947368
Booking	0.84429641965
Cancelation	0.219512195122
Rebooking	0.251184834123
Issues	0.314341846758
negotiation	0.314341846758

Q3. How could your approach be improved when e.g. spending more time or having access to the full request data?

- Would have tried with different Model - Convolutional Neural Network
- Would have tuned the hyperparameters
- Would have tried with Non-linear model (Currently I ran on linear model)
- Better feature extraction if full data was given
- Before feature extraction I would have tried using word embeddings + CNN
- Could have improved the prediction power of less-occurrence class(eg: Issues and cancellation).