Analysis:

Given dataset has reviews and ratings for hotels range from 1 to 5. After observing the dataset, one row in rating column has a text along with the number rating. So, the data needs to be cleaned, tokenized and should be converted to tensors. The converted tokens along with labels are fed into the model for training. The trained model needs to be evaluated based on given dev dataset and predictions should be made to a given test dataset.

Data exploration:

Since all the labels are balanced and no missing text or labels, we don't have to use imputing or upsampling at least for the base model.

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f2641e735f8>
```
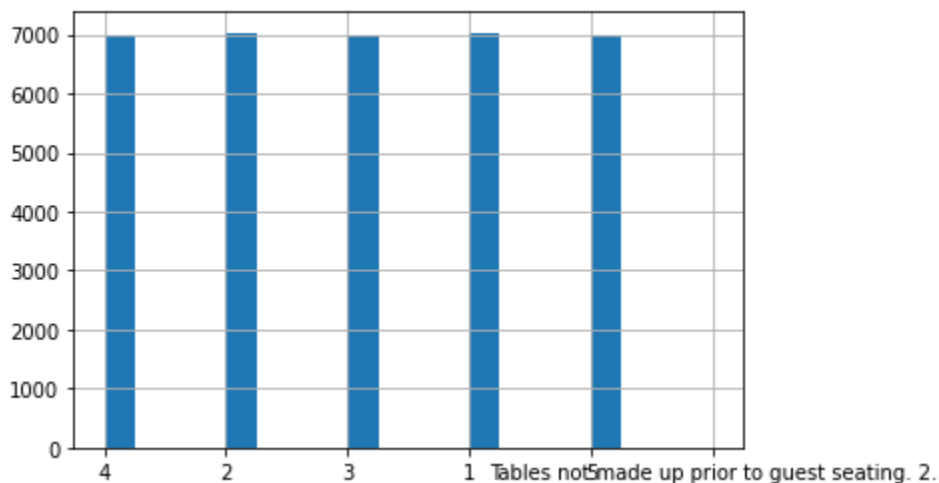

I have cleaned the rating column with regex method to select only numeric elements between 1-5.

Model Selection:

Base Model – Naïve Bayes:

Naïve Bayes is based on Bayes' theorem which assumes that each of the features it uses is conditionally independent of one another given some class.

Bayes' theorem:

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y)P(x_1, \ldots, x_n \mid y)}{P(x_1, \ldots, x_n)}$$

Using naive conditional independence, for each x_i is independent of for all i, the formula is simplified to

$$P(x_i \mid y, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = P(x_i \mid y)$$

And since $P(x_1 \ldots., x_n)$ is constant, the prediction is

$$\hat{y} = \arg\max_{y} P(y) \prod_{i=1}^{n} P(x_i \mid y)$$

The different naïve Bayes classifiers are based on assumptions they are regarding to define the distribution of $p(y/x_i)$

I have used Multinomial Naïve Bayes which is a specific instance of a Naive Bayes classifier which uses a multinomial distribution for each of the features.

TF-IDF:

Before applying Naïve Bayes, I have converted the words in sentences to tf-idf vectors. Term frequency-inverse document frequency tells us how important the word is to the document. It increases proportionally to how many times the word occurred in the document but inversely proportional to the frequency of the word in the corpus which gives less preference to commonly occurring words or stop words.

SOTA Model – LSTM (Long Short Term Memory):

Recurrent neural networks are a type of artificial neural network which is designed to recognize the pattern from the sequence of the data. It takes a fixed-size vector as input which is limited to usage of series type inputs. RNNs remembers past and its decisions are influenced by what it has learnt in the past. Basic neural networks remember things too, but they remember things they learnt during training. While RNNs learn similarly during training also besides, they remember things learnt from prior inputs while generating outputs. The hidden state captures the relationship that neighbours might have with each other in serial input and it keeps changing every time.

  LSTMs are similar to vanilla RNNs, but it introduces some changes to how we compute outputs and hidden states using inputs. LSTM introduces additional gates and cell state, which addresses the problem of keeping or resetting context across sentences regardless of the distance between such context resets. So, LSTM is a special kind of recurrent neural network that is capable of learning long-term dependencies, remembering information for long periods. A)The network decides what to learn and what to forget. B)It selectively updates cell state values. C) The network decides what part of current states makes it to output.

Evaluation measures:

Basemodel Performance:

The training accuracy of Naïve Bayes model is 78% and evaluation accuracy is 72%.

SOTA Model Performance:

The training accuracy of LSTM Model is 95% and evaluation accuracy is 73%. This accuracy is obtained for 10 epochs.

Parameters considered:

Metrics	:	accuracy
Loss function	:	categorical_crossentropy
Optimizer	:	adam
Learning rate	:	0.001 (default)
Epochs	:	7

Model comparison:

We can expect Naïve Bayes model performance is slightly less than LSTM model. This makes sense because LSTM model looks for a relationship with its neighbours within a sentence whereas Naïve Bayes calculates each word independently. Since the reviews are sometimes sarcastic, meaning though the words are positive the overall sentence is negative. Naive Bayes could not understand the sarcastic comments since it does not capture the underlying relationship which could lead to low performance.

Reference:

1. Sequence models course by Andrew NG