Data Exploration

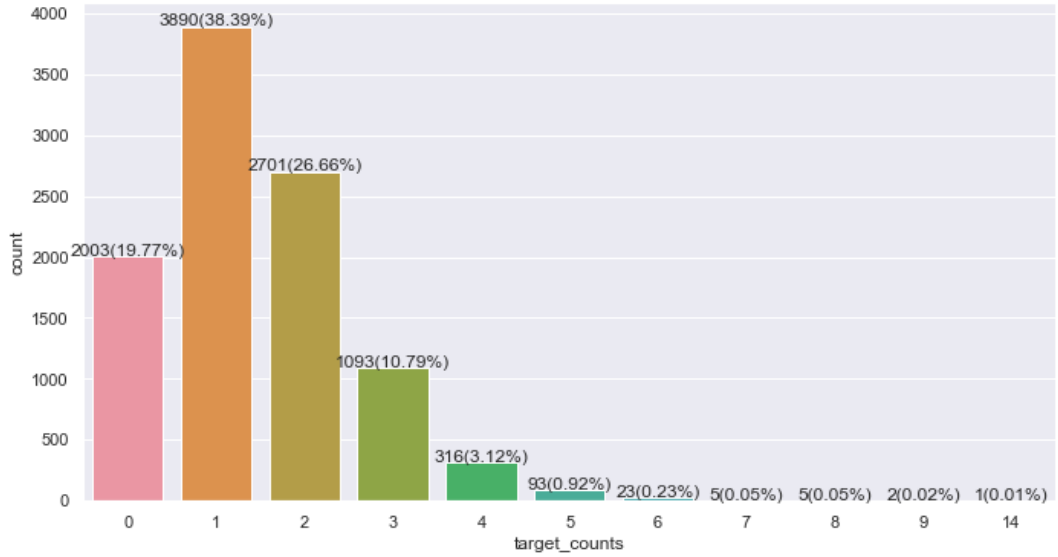
Data Loading

```
There are 10132 Rows in Data
```

Code Hidden

Distribution of Count of Labels per Datapoint

Code Hidden



There are 2003 datapoints that doesn't have any labels and most of the datapoints have less than or equal to 4 labels and one datapoint have 14 labels and that is an outlier so we will remove it

	text	target	target_counts
384	test review, should be deleted	[refund not actioned positive, refund timescal...	14

As we can see this datapoint is just for testing so we can easily drop it

```
Shape of dataframe after removing outlier datapoint (10131, 3)
```

Code Hidden

```
Number of Unique labels - 99
```

Code Hidden

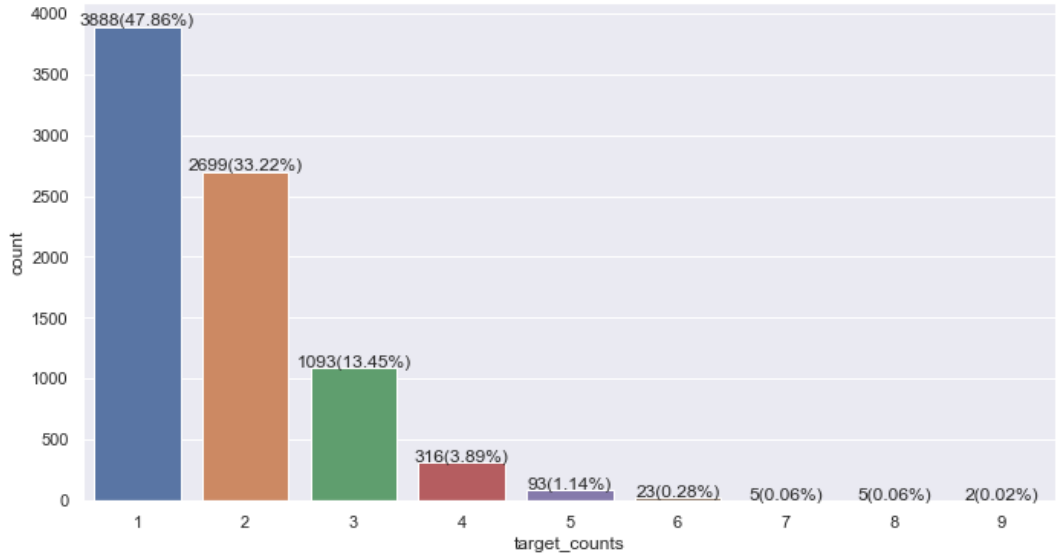
There are 99 unique labels in the dataset but alot of the labels doesn't have any sentiment

assosiated with it and also available in very low counts so we can simpy consider those as noise and can remove them.

```
Number of Noisy Labels - 44
```

Code Hidden

Code Hidden



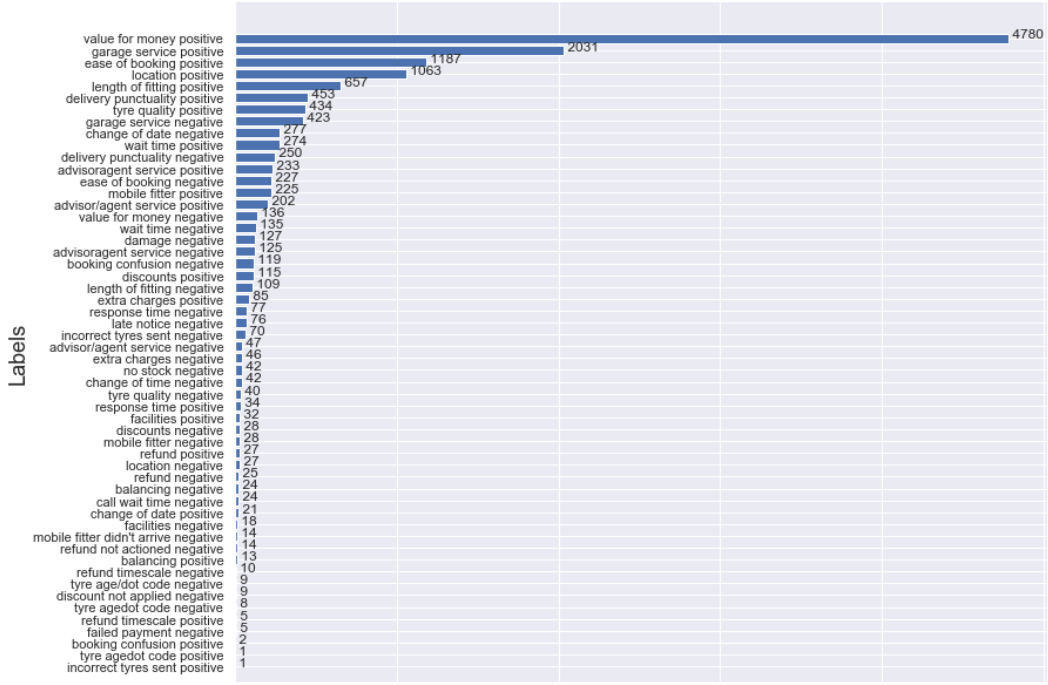
So we have removed datapoints that have 0 labels and also some noisy labels that have no sentiments assosiated with it

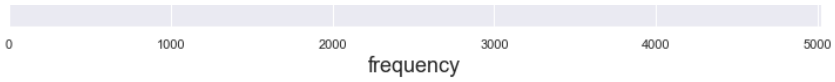
```
Number of Unique labels after removing noisy labels - 54
```

Code Hidden

After Removing noisy labels we are left with 54 labels

Count of Each Labels

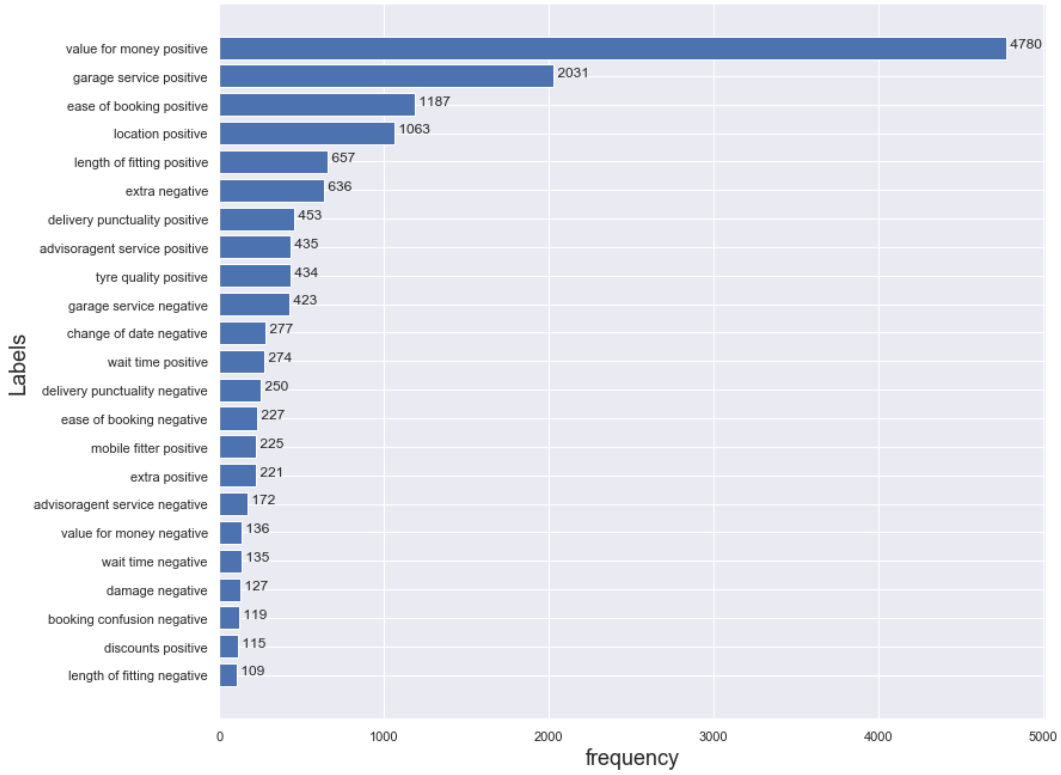




Code Hidden

As there are alot of labels that are less than 100, so we can combine them to there respective sentiment labels as due to there low frequecies it will be very difficult for model to learn and we can also there are some labels like 'advisor/agent service positive', 'advisoragent service positive' are same but due to some characters they are split so we can fix this too

Code Hidden



Code Hidden

So we have combined all the labels that have frequency less than 100 to 'extra positve' and 'extra negative' and also Labels like 'value for money positive', 'garage service positive' have very large frequency as compared to othere labels we can undersample those

```
Number of Unique Labels after all processing - 23
```

After combining less frequent labels and correcting some duplicate labels, we have 23 labels now

Under Sampling Analysis

```
No. of Rows after removing datapoint with 0 labels and noisy labels - 8124
```

Code Hidden

Undersampling Analysis for value for money positive label

```
No. of rows with only 'value for money positive' as Label - 2143
```

Code Hidden

There are 2143 datapoints in which there is only one label that is value for money positive, we can directly undersample these

```
No. of rows after undersampling 'value for money positive' - 6195
```

Code Hidden

After Removing 90% of single occuring value for money positive label we have 6195 datapoints

Undersampling Analysis for garage service positive label

```
No. of rows with only 'garage service positive' as Label - 558
```

Code Hidden

There are 558 datapoints in which there is only one label that is 'garage service positive', we can directly undersample these

```
No. of rows after undersampling 'garage service positive' - 5749
```

Code Hidden

After Removing 80% of single occuring garage service positive label we have 5749 datapoints

Undersampling Analysis for value for money positive, garage service positive label together

```
No. of rows with both 'value for money positive' and 'garage service positive'
as Label - 232
```

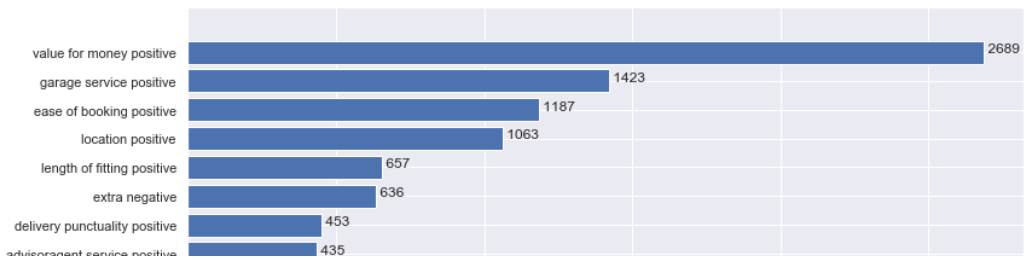
Code Hidden

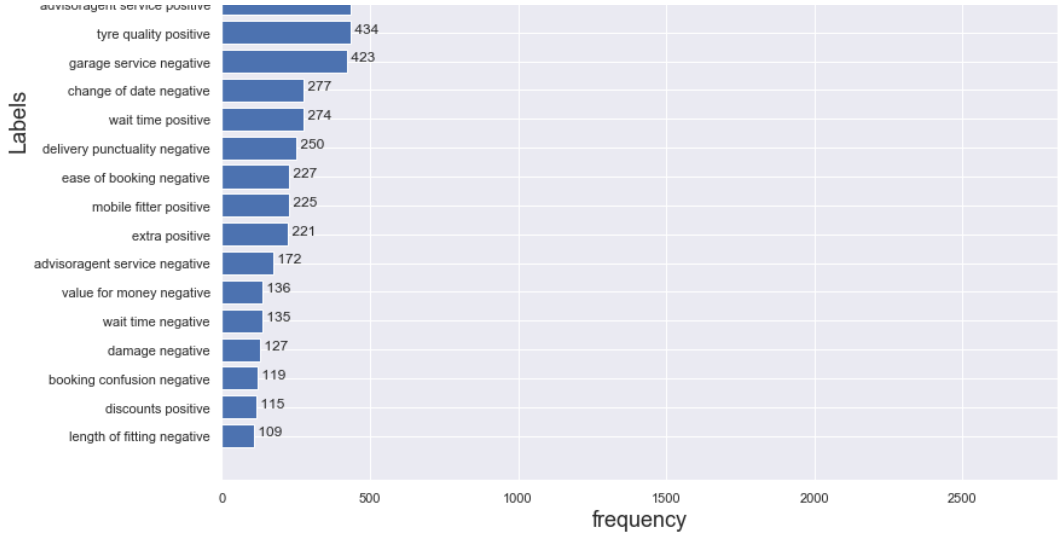
```
No. of rows after undersampling 'value for money positive' and 'garage service
positive' together - 5587
```

Code Hidden

After Undersampling of most frequent occuring labels we have 5587 datapoints

Final Distribution of Labels

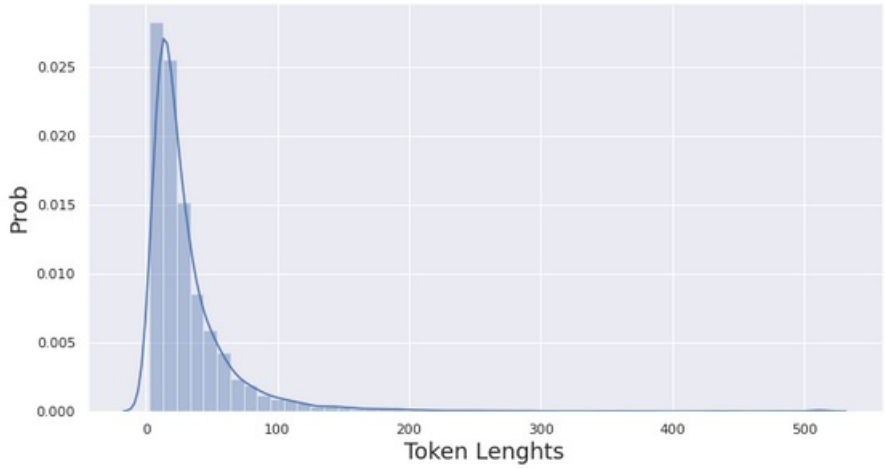




Code Hidden

We have reduced the most frequently occuring labels significantly

Distibution of Token Lengths of Texts



Most of the tokens length will come under 128