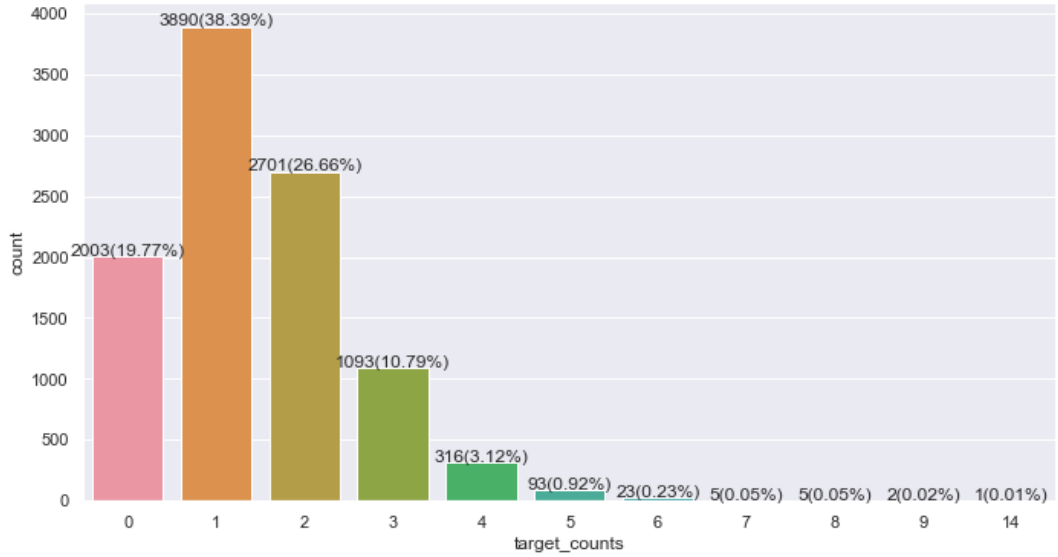
Data Exploration

Data Loading

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

Distribution of Count of Labels per Datapoint



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)
```

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

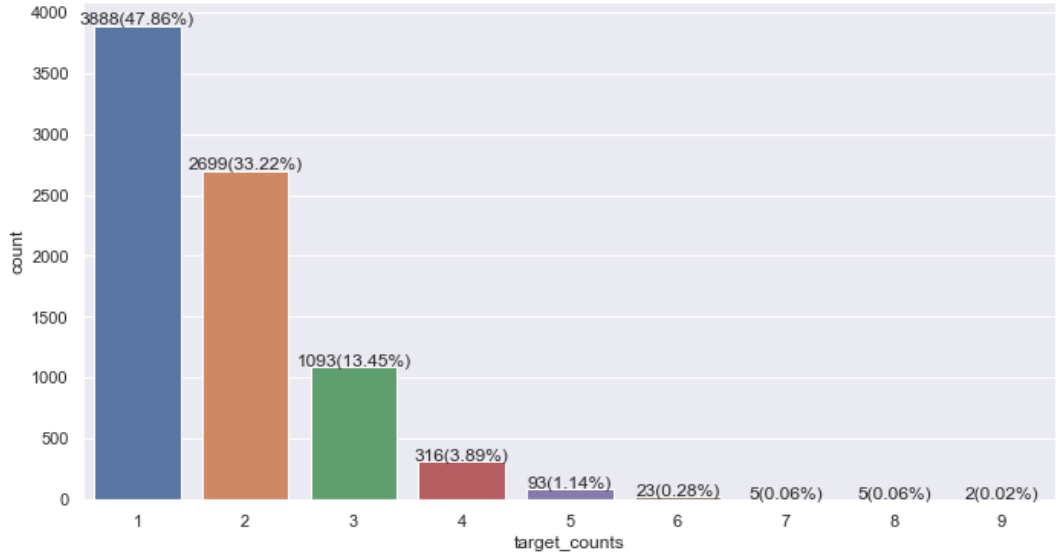
	Count
value for money positive	4780
garage service positive	2031
ease of booking positive	1187
location positive	1063
length of fitting positive	657
delivery punctuality positive	453
tyre quality positive	434

	Count
garage service negative	423
change of date negative	277
wait time positive	274
delivery punctuality negative	250
advisoragent service positive	233
ease of booking negative	227
mobile fitter positive	225
advisor/agent service positive	202
value for money negative	136
wait time negative	135
damage negative	127
advisoragent service negative	125
booking confusion negative	119
discounts positive	115
length of fitting negative	109
extra charges positive	85
response time negative	77
late notice negative	76
incorrect tyres sent negative	70
advisor/agent service negative	47
extra charges negative	46
no stock negative	42
change of time negative	42
tyre quality negative	40
response time positive	34
facilities positive	32
mobile fitter negative	28
discounts negative	28
refund positive	27
location negative	27
refund negative	25
call wait time negative	24
balancing negative	24
change of date positive	21
facilities negative	18
refund not actioned negative	14
mobile fitter didn't arrive negative	14
balancing positive	13
refund timescale negative	10
tyre age/dot code negative	9
discount not applied negative	9
tyre agedot code negative	8
refund timescale positive	5
failed payment negative	5
good price	2
booking confusion positive	2
good communication	1
really good prices	1
simple fitting procedure	1
good prices	1

	Count
telephone help avaiable if required.	1
it feels more transparent than buying from a mechanic based on their opinion.	1
easily navigable web site	1
faster than dealerships to arrange supply and partnership with ATS works well .	1
tyre agedot code positive	1
staff were courteous	1
ability to browse through various makes of tyres before making a decision	1
)	1
incorrect tyres sent positive	1
recommended a great local shop for fitting	1
wide choice of tyres at great prices including delivery and fitting.	1
no issues	1
keep it up! Will definitely buy from you again.	1
my tyre supplier of choice!.	1
one tyre with dangerously low air pressure. Asked the garage to top up the tyre and check the other three. They did top up the one tyre but didn't bother to check the others. When I checked the tyres 10 minutes later all four tyres had varying pressures (from 29 to 37.5). I emailed both companies and have not received an apology from either. I will never use Lavender Motors again (they shouldn't be in business) and likely not [REDACTED].com either.	1
won’t be using anyone else.	1
clean reception area and free coffee	1
only 8 as the garage was unable to fit the tyres in the allotted time and I had to wait over an hour.	1
great price.	1
simple to use website	1
and an excellent service from the garage who fitted the tyre.	1
cheaper than anything else I have found by some way.	1
polite staff	1
prompt deliveries...	1
good range and competitive prices on website	1
fitting took a long time.	1
the process was simple and quick.Regular offers make it easy to find premium tyres at decent prices.	1
slick delivery service and good fitter.	1
good local fitting services	1
and both times have been good experiences.	1
Great garage fitted them	1
Cheapest price	1
Garage was quick & efficient with fitting of tyres	1
they looked grey and old (but unused) not at all how new tyres normally look. Not surprising when I inspected the 'Dot Date' and it was 4517 (ie tyres manufactured in week 45 of 2017.Queried this with [REDACTED] who told me that they have many different suppliers and do not check the age of the tyres.They could re-order but no guarantee that the replacements would be any newer. Not at all happy and requested a refund which they say they will do.	1
the site is easy to use	1
simple booking for fitting procedure.	1
good service	1
easy to browse selection of tyres before making a choice	1
tje garage i choice to go to in order to have the tyres fitted was nit that good and i certainly will not be using them again.	1
really friendly and efficient service.	1
hassle free	1

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
```

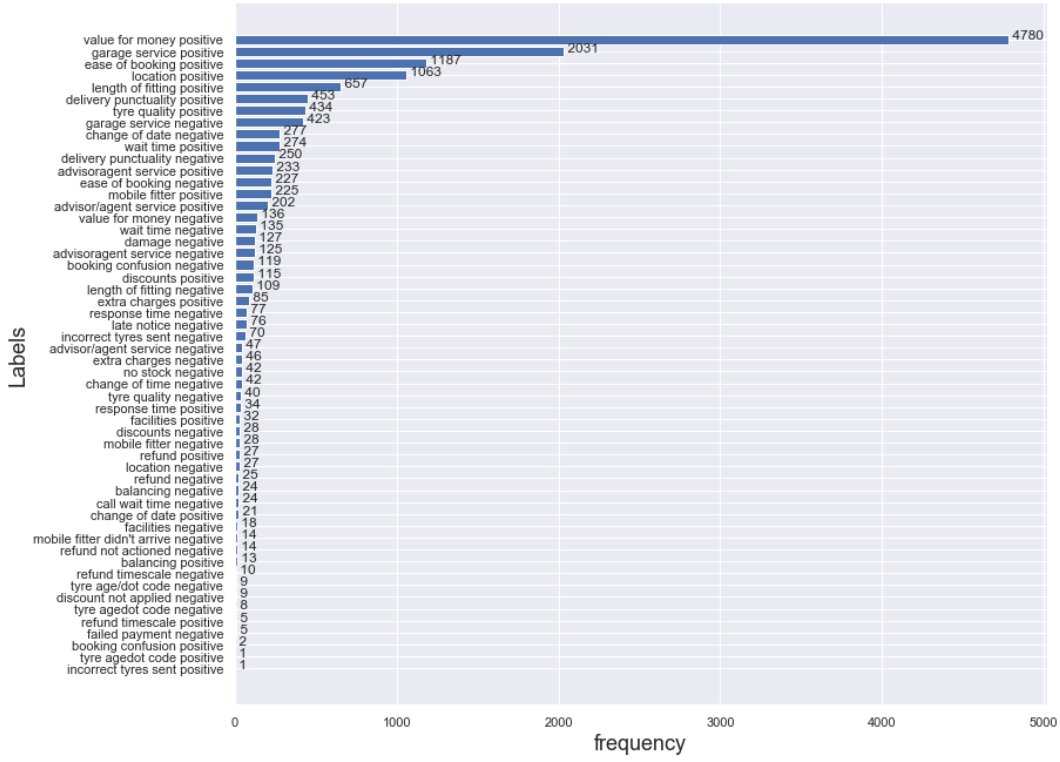


So we have removed datapoints that have 0 labels and also some noisy labels that have no sentiments assosiated with it

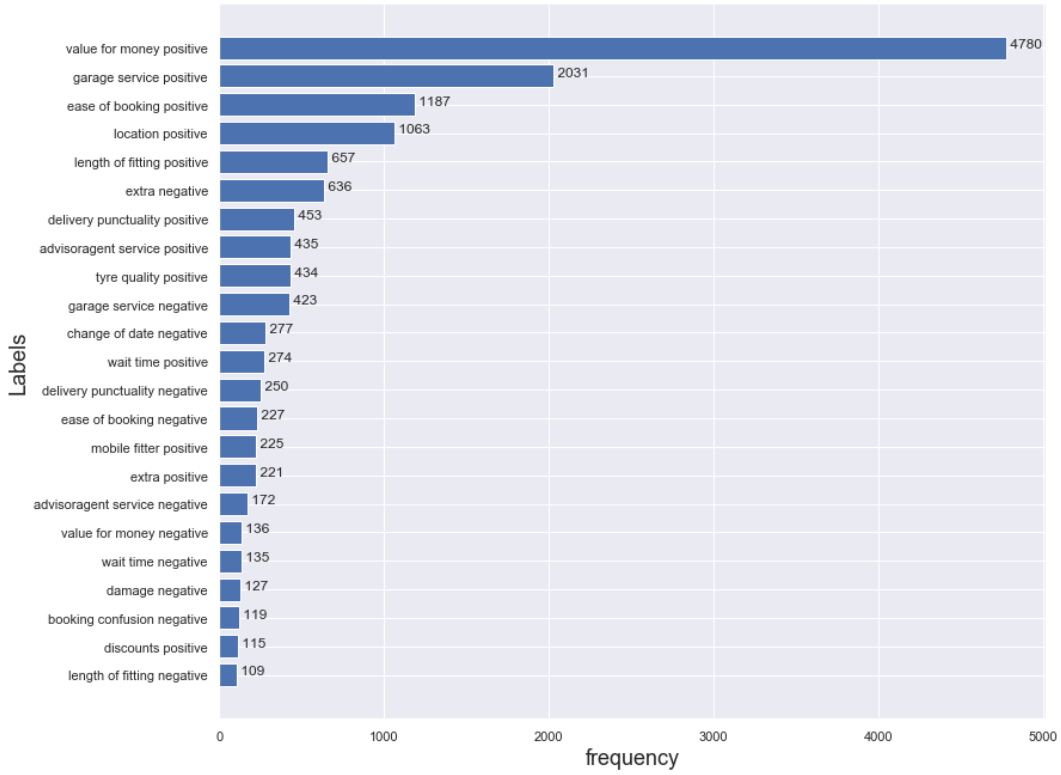
```
Number of Unique labels after removing noisy labels - 54
```

After Removing noisy labels we are left with 54 labels

Count of Each Labels



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



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
```

Undersampling Analysis for value for money positive label

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

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
```

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
```

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
```

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

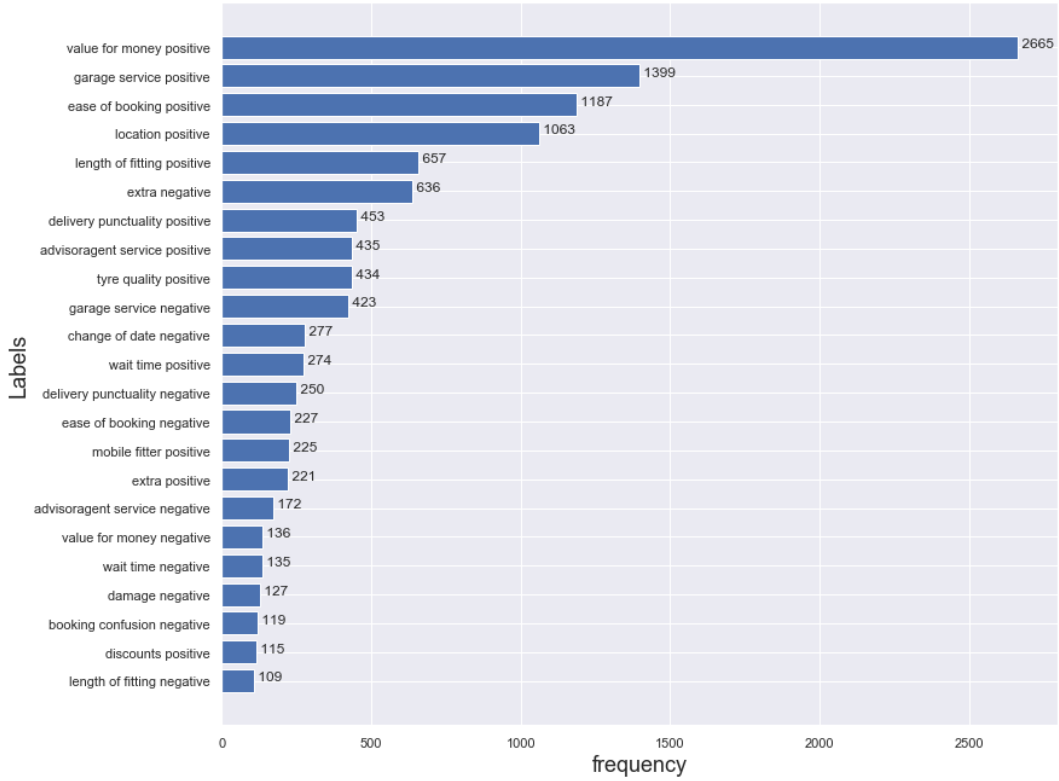
Undersampling Analysis for value for money positive, garage service positive label together

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

```
No. of rows after undersampling 'value for money positive' and 'garage service
positive' together - 5563
```

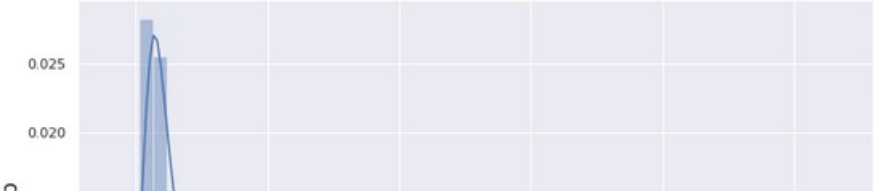
After Undersampling of most frequent occuring labels we have 5587 datapoints

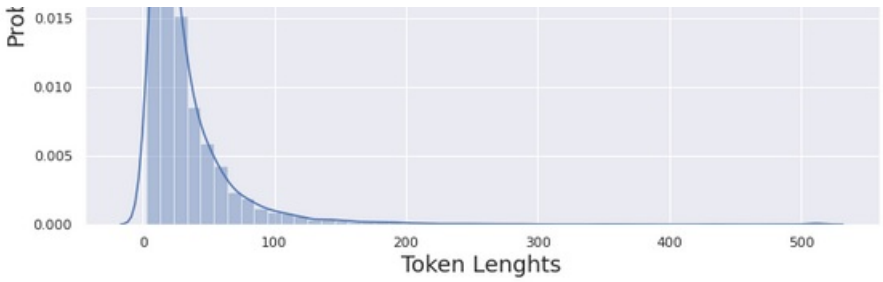
Final Distribution of Labels



We have reduced the most frequently occuring labels significantly

Distibution of Token Lengths of Texts





Most of the tokens length will come under 128