[nltk\_data] Downloading package punkt to /root/nltk\_data...

[nltk\_data] Package punkt is already up-to-date!

[nltk\_data] Downloading package stopwords to /root/nltk\_data...

[nltk\_data] Package stopwords is already up-to-date!

[nltk\_data] Downloading package omw-1.4 to /root/nltk\_data...

[nltk\_data] Package omw-1.4 is already up-to-date!

DATAFRAME contains how many rows and columns:

(3306, 6)

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 3306 entries, 0 to 3305

Data columns (total 6 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Unnamed: 0 3306 non-null int64

1 Restaurant\_Name 3306 non-null object

2 Restaurant\_rated 3306 non-null float64

3 Review\_Written\_By 3306 non-null object

4 User\_Rated\_Restaurant 3306 non-null float64

5 Written\_Review 1909 non-null object

dtypes: float64(2), int64(1), object(3)

memory usage: 155.1+ KB

DATATYPES:

FIRST ROWS OF DATAFRAME ARE:

Unnamed: 0 Restaurant\_Name Restaurant\_rated Review\_Written\_By \

0 0 Toit 4.7 Brijmohan Patil

1 1 Toit 4.7 Jasleen Rajpal

2 2 Toit 4.7 Rukmini Das

3 3 Toit 4.7 Ritu Upadhyay

4 4 Toit 4.7 Sambhavi

User\_Rated\_Restaurant Written\_Review

0 4.0 NaN

1 5.0 Best food ever! And the packaging was a solid ...

2 4.0 crumb could be a litter lesser n fish more,non...

3 1.0 NaN

4 5.0 NaN

How many NULLS are their in zomato Dataframe:

Unnamed: 0 0

Restaurant\_Name 0

Restaurant\_rated 0

Review\_Written\_By 0

User\_Rated\_Restaurant 0

Written\_Review 1397

dtype: int64

Unnamed: 0 Restaurant\_Name Restaurant\_rated Review\_Written\_By \

0 0 Toit 4.7 Brijmohan Patil

1 1 Toit 4.7 Jasleen Rajpal

2 2 Toit 4.7 Rukmini Das

3 3 Toit 4.7 Ritu Upadhyay

4 4 Toit 4.7 Sambhavi

... ... ... ... ...

3301 3301 Cafe Paashh 4.9 Shuchi Saxena

3302 3302 Cafe Paashh 4.9 Mohnish Chordiya

3303 3303 Cafe Paashh 4.9 Aditya Shrimankar

3304 3304 Cafe Paashh 4.9 Nehaa T.

3305 3305 Cafe Paashh 4.9 Dishita Shah

User\_Rated\_Restaurant Written\_Review

0 4.0 -

1 5.0 Best food ever! And the packaging was a solid ...

2 4.0 crumb could be a litter lesser n fish more,non...

3 1.0 -

4 5.0 -

... ... ...

3301 5.0 -

3302 5.0 -

3303 3.0 -

3304 3.0 not too good. was expecting more from here sin...

3305 5.0 -

[3306 rows x 6 columns]

Splitted Review By Spaces:

0 1

1 11

2 16

3 1

4 1

..

3301 1

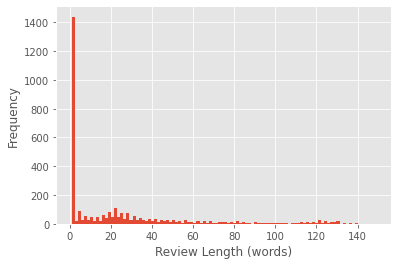
3302 1

3303 1

3304 13

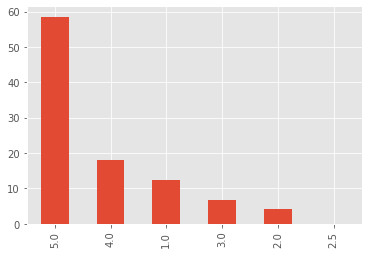
3305 1

Name: Written\_Review, Length: 3306, dtype: int64



Average words: 25.171203871748336

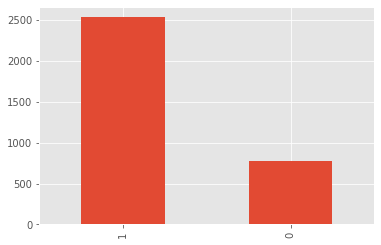
Skewness: 1.6621400523063772



1 2535

0 771

Name: Sentiment\_Rating, dtype: int64



After EDA dataframe has rows and columns: (3306, 7)

Number of unique tokens then: 7848

Number of unique tokens then: 6477

-Old Review:-

looks good, delivery boy spoiled it

-New Review-

looks good delivery boy spoiled it

Number of unique tokens then: 7848

Number of unique tokens then: 5770

['arabic', 'azerbaijani', 'basque', 'bengali', 'catalan', 'chinese', 'danish', 'dutch', 'english', 'finnish', 'french', 'german', 'greek', 'hebrew', 'hinglish', 'hungarian', 'indonesian', 'italian', 'kazakh', 'nepali', 'norwegian', 'portuguese', 'romanian', 'russian', 'slovene', 'spanish', 'swedish', 'tajik', 'turkish']

Original sentence best food ever and the packaging was a solid 10 10 loved it

Stopwords in the sentence ['and', 'the', 'was', 'a', 'it']

Non-stopwords in the sentence ['best', 'food', 'ever', 'packaging', 'solid', '10', '10', 'loved']

Written\_Review\_new \

0

1 best food ever and the packaging was a solid ...

2 crumb could be a litter lesser n fish more non...

3

4

... ...

3301

3302

3303

3304 not too good was expecting more from here sin...

3305

Written\_Review\_nonstop

0 []

1 [best, food, ever, packaging, solid, 10, 10, l...

2 [crumb, could, litter, lesser, n, fish, noneth...

3 []

4 []

... ...

3301 []

3302 []

3303 []

3304 [good, expecting, since, prices, high]

3305 []

[3306 rows x 2 columns]

[nltk\_data] Downloading package wordnet to /root/nltk\_data...

[nltk\_data] Package wordnet is already up-to-date!

Written\_Review\_nonstop \

0 []

1 [best, food, ever, packaging, solid, 10, 10, l...

2 [crumb, could, litter, lesser, n, fish, noneth...

3 []

4 []

... ...

3301 []

3302 []

3303 []

3304 [good, expecting, since, prices, high]

3305 []

Written\_Review\_stemm \

0 []

1 [best, food, ev, pack, solid, 10, 10, lov]

2 [crumb, could, lit, less, n, fish, nonetheless...

3 []

4 []

... ...

3301 []

3302 []

3303 []

3304 [good, expect, sint, pric, high]

3305 []

Written\_Review\_lemm

0 []

1 [best, food, ever, packaging, solid, 10, 10, l...

2 [crumb, could, litter, lesser, n, fish, noneth...

3 []

4 []

... ...

3301 []

3302 []

3303 []

3304 [good, expecting, since, price, high]

3305 []

[3306 rows x 3 columns]

00 000 000rs 00pm 10 100 1000 100rs 1058 10mins ... yumm \

0 0 0 0 0 0 0 0 0 0 0 ... 0

1 0 0 0 0 2 0 0 0 0 0 ... 0

2 0 0 0 0 0 0 0 0 0 0 ... 0

3 0 0 0 0 0 0 0 0 0 0 ... 0

4 0 0 0 0 0 0 0 0 0 0 ... 0

yummiest yummmm yummy zafrani zero zest zishan zomato zone

0 0 0 0 0 0 0 0 0 0

1 0 0 0 0 0 0 0 0 0

2 0 0 0 0 0 0 0 0 0

3 0 0 0 0 0 0 0 0 0

4 0 0 0 0 0 0 0 0 0

[5 rows x 5739 columns]

(1, 627) 1

(1, 2049) 1

(1, 1814) 1

(1, 319) 1

(1, 5152) 1

(1, 3613) 1

(1, 5566) 1

(1, 133) 1

(1, 4719) 1

(1, 6) 2

(1, 3026) 1

(1, 2672) 1

(2, 319) 1

(2, 5152) 1

(2, 133) 1

(2, 1354) 1

(2, 1287) 1

(2, 575) 1

(2, 2967) 1

(2, 2912) 1

(2, 3344) 1

(2, 2006) 1

(2, 3281) 1

(2, 3423) 1

(2, 5144) 1

: :

(3299, 5518) 1

(3299, 3336) 1

(3299, 3187) 1

(3299, 454) 1

(3299, 2675) 1

(3299, 3284) 1

(3299, 5461) 1

(3299, 2480) 1

(3299, 2368) 1

(3299, 867) 1

(3299, 3717) 1

(3299, 1118) 1

(3304, 5566) 1

(3304, 3281) 1

(3304, 2220) 1

(3304, 3435) 1

(3304, 390) 1

(3304, 2407) 1

(3304, 4708) 1

(3304, 2414) 1

(3304, 2125) 1

(3304, 5256) 1

(3304, 3949) 1

(3304, 4632) 1

(3304, 1869) 1

1 0.787009

0 0.212991

Name: Sentiment\_Rating, dtype: float64

MultinomialNB()

Prediction:

[1 1 0 1 1 1 1 1 1 1 0 1 1 1 1 1 0 1 1 1 1 0 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1

1 1 1 1 1 1 1 1 1 0 1 1 0 0 1 0 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1

1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 0 0 1 1 1 0 1 1 1 1 1 1 1 1

1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 0 0 1 1 0 1 1 1 1 1 0 1 1 1 1 1 1 0 0

1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 1 1 1 1 0 0 1 1 1 0 1 1 1 1 1 1 1

1 1 1 1 1 1 1 1 1 0 1 1 0 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 1 1

1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 0 1 1 0 1 1 1 1 1 1

1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1

1 1 1 1 1 1 0 1 1 1 0 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 0 1 1 1 1 0 1 1 0 1

1 1 1 0 1 1 1 1 1 1 0 1 1 1 1 1 0 1 1 1 1 0 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1

0 1 1 0 1 1 1 1 1 0 1 0 1 0 1 1 1 0 0 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 0 1 1

1 1 1 1 1 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 0 1 1 0 1 1 1 1 0 0 1

1 1 0 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 0 1 1 1 1 1 1 0 1 1 1 0 1 1 1

1 1 1 1 1 1 0 0 1 0 1 1 1 0 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

1 1 1 1 1 1 0 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1

1 1 1 0 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 0 0 1

1 0 1 0 1 1 1 0 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 0 1 1

1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 0 1 1 1 1 1 1 0 1 0]

Actual:

587 1

33 1

2180 0

2155 1

1779 1

..

2567 1

3210 1

1114 0

2979 1

2896 0

Name: Sentiment\_Rating, Length: 662, dtype: int64

CONFUSION MATRIX WITH BAG-OF-WORDS: [[ 78 15]

[ 63 506]]

OUR MODEL HAS ACCURACY WITH BAG-OF-WORDS: 88.21752265861026 %

F1 score: 0.9184234647112739

F1 score: 0.9197080291970803

words weights

66432 great 1.295028

10034 amazing 1.137025

125940 rohit 1.033489

52032 excellent 0.971883

23684 awesome 0.927983

28713 best 0.829009

99740 nice 0.791629

64075 good 0.774817

90290 love 0.566638

11172 ambience 0.548218

90525 loved 0.516927

44537 delicious 0.461252

167105 vibe 0.459149

160341 toit 0.437297

115274 place 0.431114

words weights

126300 rubbish -0.552890

142053 taste -0.611075

153069 they -0.612738

32525 but -0.684858

111226 overrated -0.700857

112709 pathetic -0.702775

52588 expensive -0.706788

87641 less -0.819078

23392 average -0.842831

24295 bad -0.905755

108609 order -0.941093

100513 no -1.013946

39382 cold -1.068418

101158 not -1.350875

181286 worst -1.622384

MultinomialNB()

Prediction:

[1 1 0 1 0 1 1 1 1 1 0 1 1 1 1 1 0 1 1 0 1 0 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1

1 1 1 1 1 1 1 1 1 0 1 1 0 0 1 0 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 0 1 1 1 1 1

1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 0 0 1 1 1 1 1 1 1 1 0 1 1 1

1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 0 0 1 1 0 1 1 1 1 1 0 1 1 1 1 1 1 0 1

1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 0 1 0 0 1 1 1 1 0 0 1 1 1 0 1 1 1 1 1 1 1

1 1 1 1 1 1 1 1 1 0 1 1 0 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 1 1

1 1 1 1 0 1 1 1 1 0 1 1 1 1 1 1 1 0 1 0 1 1 1 1 1 1 1 0 1 1 0 1 1 1 1 1 1

1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 0 1 1

1 1 1 1 1 1 0 1 1 1 0 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 0 1 1 1 1 0 1 1 0 1

1 1 1 0 1 1 1 1 1 1 0 1 1 1 1 1 0 1 1 1 1 0 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1

0 1 1 0 1 1 1 1 1 0 1 0 1 0 1 1 1 0 0 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 0 1 1

1 1 1 1 1 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 0 1 1 0 1 1 1 1 0 0 1

1 1 0 1 1 1 0 1 1 1 1 1 1 1 1 1 1 0 1 1 0 1 0 1 1 1 1 1 1 0 1 1 1 0 1 1 1

1 1 1 1 1 0 0 0 1 0 1 1 1 0 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

1 1 1 1 1 1 0 1 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1

1 1 1 0 1 1 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 0 1 0 1 1

1 0 1 0 1 1 1 0 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 0 1 1

1 1 1 1 1 1 1 0 1 0 1 1 1 1 1 1 0 1 1 1 1 1 1 0 1 1 1 1 1 1 0 1 0]

Actual:

587 1

33 1

2180 0

2155 1

1779 1

..

2567 1

3210 1

1114 0

2979 1

2896 0

Name: Sentiment\_Rating, Length: 662, dtype: int64

CONFUSION MATRIX WITH TD-IDF: [[ 84 23]

[ 57 498]]

OUR MODEL HAS ACCURACY WITH TD-IDF: 87.91540785498489 %

F1 score TD-IDF: 0.9201793721973094