Predict_Sentiments_of_Amazon_Customer

August 3, 2023

1 Predict Sentiments of Amazon Customer

1.1 Q.Preprocessing

[1]:
```
# libraries
import pandas as pd          # for data manipulation and data analysis
import numpy as np           # for large and multi dimensional array
```

[2]:
```
# load data
data = pd.read_csv('Reviews.csv')
data
```

[2]:

	Id	ProductId	UserId	ProfileName	\
0	1	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	
1	2	B00813GRG4	A1D87F6ZCVE5NK	dll pa	
2	3	B000LQOCH0	ABXLMWJIXXAIN	Natalia Corres "Natalia Corres"	
3	4	B000UA0QIQ	A395BORC6FGVXV	Karl	
4	5	B006K2ZZ7K	A1UQRSCLF8GW1T	Michael D. Bigham "M. Wassir"	
…	…	…	…	…	
568449	568450	B001EO7N10	A28KG5XORO54AY	Lettie D. Carter	
568450	568451	B003S1WTCU	A3I8AFVPEE8KI5	R. Sawyer	
568451	568452	B004I613EE	A121AA1GQV751Z	pksd "pk_007"	
568452	568453	B004I613EE	A3IBEVCTXKNOH	Kathy A. Welch "katwel"	
568453	568454	B001LR2CU2	A3LGQPJCZVL9UC	srfell17	

	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	\
0	1	1	5	1303862400	
1	0	0	1	1346976000	
2	1	1	4	1219017600	
3	3	3	2	1307923200	
4	0	0	5	1350777600	
…	…	…	…	…	
568449	0	0	5	1299628800	
568450	0	0	2	1331251200	
568451	2	2	5	1329782400	
568452	1	1	5	1331596800	
568453	0	0	5	1338422400	

```
                                Summary  \
0                   Good Quality Dog Food
1                       Not as Advertised
2                   "Delight" says it all
3                          Cough Medicine
4                             Great taffy
…                                     …
568449                Will not do without
568450                      disappointed
568451           Perfect for our maltipoo
568452  Favorite Training and reward treat
568453                        Great Honey

                                                     Text
0       I have bought several of the Vitality canned d…
1       Product arrived labeled as Jumbo Salted Peanut…
2       This is a confection that has been around a fe…
3       If you are looking for the secret ingredient i…
4       Great taffy at a great price.  There was a wid…
…                                                     …
568449  Great for sesame chicken..this is a good if no…
568450  I'm disappointed with the flavor. The chocolat…
568451  These stars are small, so you can give 10-15 o…
568452  These are the BEST treats for training and rew…
568453  I am very satisfied ,product is as advertised,…

[568454 rows x 10 columns]
```

[3]:
```
print(data.columns)
print(data.size)
print(data.dtypes)
print(data.shape)
```

```
Index(['Id', 'ProductId', 'UserId', 'ProfileName', 'HelpfulnessNumerator',
       'HelpfulnessDenominator', 'Score', 'Time', 'Summary', 'Text'],
      dtype='object')
5684540
Id                          int64
ProductId                  object
UserId                     object
ProfileName                object
HelpfulnessNumerator        int64
HelpfulnessDenominator      int64
Score                       int64
Time                        int64
Summary                    object
Text                       object
dtype: object
```

(568454, 10)

1.2 Data Preparation

[4]:
```
data['Helpful%'] = np.where(data['HelpfulnessDenominator']>0,␣
 ↪data['HelpfulnessNumerator']/data['HelpfulnessDenominator'], -1)
data['Helpful%']
```

[4]:
```
0         1.0
1        -1.0
2         1.0
3         1.0
4        -1.0
         ...
568449   -1.0
568450   -1.0
568451    1.0
568452    1.0
568453   -1.0
Name: Helpful%, Length: 568454, dtype: float64
```

add different label according to the values

[5]:
```
data['Helpful%'].unique()
```

[5]:
```
array([ 1.        , -1.        ,  0.8       ,  0.        ,  0.5       ,
        0.66666667,  0.25      ,  0.89473684,  0.83333333,  0.75      ,
        0.33333333,  0.3       ,  0.11111111,  0.42857143,  0.875     ,
        0.85714286,  0.2       ,  0.26315789,  0.6       ,  0.71428571,
        0.53846154,  0.57142857,  0.91489362,  0.86666667,  0.82352941,
        0.78571429,  0.74074074,  0.4       ,  0.375     ,  0.28571429,
        0.14285714,  0.77777778,  0.125     ,  0.9       ,  0.94117647,
        0.92307692,  0.7       ,  0.45454545,  0.88888889,  0.83870968,
        0.9047619 ,  0.92857143,  0.90909091,  0.91666667,  0.84615385,
        0.10526316,  0.98214286,  0.97826087,  0.7518797 ,  0.3125    ,
        0.1       ,  0.18518519,  0.88      ,  0.69230769,  0.625     ,
        0.54545455,  0.41666667,  0.45833333,  0.22222222,  0.81818182,
        0.8125    ,  0.16666667,  0.93103448,  0.88235294,  0.23529412,
        0.63636364,  0.81481481,  0.95652174,  0.64285714,  0.58333333,
        0.94444444,  0.921875  ,  0.86574074,  0.96      ,  0.91304348,
        0.64705882,  0.95833333,  0.09090909,  0.13333333,  0.52941176,
        0.96969697,  0.36363636,  0.07142857,  0.72727273,  0.18181818,
        0.96666667,  0.99074074,  0.97297297,  0.80645161,  0.64102564,
        0.55555556,  0.4375    ,  0.76923077,  0.28      ,  0.15384615,
        0.44444444,  0.5625    ,  0.53333333,  0.47058824,  0.47222222,
        0.23076923,  0.25925926,  0.98876404,  0.88372093,  0.19047619,
        0.94594595,  0.84313725,  0.96629213,  0.72222222,  0.05882353,
        0.27272727,  0.97959184,  0.26666667,  0.30769231,  0.94736842,
```

```
0.27777778,  0.6875    ,  0.92      ,  0.90566038,  0.95      ,
0.9375    ,  0.9137931 ,  0.82857143,  0.86363636,  0.85      ,
0.96428571,  0.95238095,  0.08333333,  0.97560976,  0.93333333,
0.46666667,  0.96153846,  0.24      ,  0.92682927,  0.93548387,
0.86956522,  0.06666667,  0.98461538,  0.97      ,  0.97619048,
0.925     ,  0.88461538,  0.61538462,  0.09375   ,  0.79166667,
0.70588235,  0.45      ,  0.93939394,  0.90322581,  0.68      ,
0.95454545,  0.04166667,  0.89655172,  0.88571429,  0.38461538,
0.07692308,  0.12121212,  0.92237443,  0.92156863,  0.36585366,
0.88095238,  0.84      ,  0.61904762,  0.96129032,  0.96385542,
0.90588235,  0.87878788,  0.05555556,  0.80952381,  0.20689655,
0.07407407,  0.35      ,  0.77272727,  0.91428571,  0.04545455,
0.76470588,  0.70833333,  0.73333333,  0.93650794,  0.8671875 ,
0.75949367,  0.65957447,  0.57692308,  0.41176471,  0.40909091,
0.34693878,  0.30263158,  0.16176471,  0.65      ,  0.96296296,
0.96808511,  0.94915254,  0.98290598,  0.9893617 ,  0.95744681,
0.96268657,  0.98305085,  0.61111111,  0.59183673,  0.98913043,
0.98809524,  0.92982456,  0.78947368,  0.75757576,  0.82608696,
0.96491228,  0.84507042,  0.98412698,  0.96551724,  0.87341772,
0.73913043,  0.7037037 ,  0.98888889,  0.7826087 ,  0.17647059,
0.96226415,  0.94339623,  0.97058824,  0.57894737,  0.47368421,
0.5106383 ,  0.97777778,  0.92352941,  0.78378378,  0.97674419,
0.35714286,  0.94805195,  0.94285714,  0.86538462,  0.43478261,
0.99186992,  0.8627451 ,  0.97142857,  0.98484848,  0.73076923,
0.68181818,  0.63333333,  0.64583333,  0.96774194,  0.05263158,
0.36842105,  0.82926829,  0.92045455,  0.34782609,  0.85365854,
0.91803279,  0.97222222,  0.46153846,  0.2173913 ,  0.82051282,
0.29032258,  0.95754717,  0.91176471,  0.04761905,  0.65714286,
0.13636364,  0.77142857,  0.953125  ,  0.92592593,  0.0862069 ,
0.80555556,  0.20512821,  0.29411765,  0.9925187 ,  0.98564593,
0.99253731,  0.80487805,  0.82142857,  0.76      ,  0.21428571,
0.31914894,  0.02702703,  0.20833333,  0.92105263,  0.78125   ,
0.61290323,  0.97435897,  0.07894737,  0.72413793,  0.03125   ,
0.68421053,  0.97979798,  0.38888889,  0.975     ,  0.80769231,
0.06060606,  0.93023256,  0.97260274,  0.90769231,  0.31372549,
0.15789474,  0.32258065,  0.95959596,  0.21052632,  0.84210526,
0.32      ,  0.92631579,  0.03703704,  1.5       ,  0.11428571,
0.88333333,  0.1875    ,  0.96875   ,  0.64      ,  0.30434783,
0.93150685,  0.88709677,  0.75609756,  0.60606061,  0.54166667,
0.52380952,  0.98275862,  0.98630137,  0.76190476,  0.85106383,
0.79069767,  0.8974359 ,  0.93617021,  0.87234043,  0.0625    ,
0.075     ,  0.39393939,  0.74107143,  0.49090909,  0.90243902,
0.56521739,  0.27027027,  0.03846154,  0.31147541,  0.24528302,
0.97727273,  0.60714286,  0.98360656,  0.95918367,  0.94      ,
0.72      ,  0.15      ,  0.12903226,  0.35294118,  0.14084507,
0.13888889,  0.08219178,  0.03636364,  0.13043478,  0.55172414,
0.64516129,  0.98      ,  0.76271186,  0.98333333,  0.95384615,
```

```
0.85294118,  0.13513514,  0.32142857,  0.87912088,  0.82758621,
0.72881356,  0.73684211,  0.86111111,  0.81355932,  0.72839506,
0.73809524,  0.74193548,  0.51612903,  0.7109375 ,  0.69565217,
0.02941176,  0.17391304,  0.85185185,  0.06451613,  0.92727273,
0.08695652,  0.03333333,  0.475     ,  0.32352941,  0.22727273,
0.98113208,  0.42307692,  3.        ,  0.90625   ,  0.8404908 ,
0.72093023,  0.98181818,  0.69047619,  0.05660377,  0.93159609,
0.95604396,  0.95348837,  0.98823529,  0.95774648,  0.94520548,
0.62068966,  0.22058824,  0.25827815,  0.86842105,  0.82222222,
0.89041096,  0.78846154,  0.63157895,  0.98717949,  0.93406593,
0.11538462,  0.04      ,  0.86206897,  0.38095238,  0.95555556,
0.97402597,  0.94230769,  0.47619048,  0.99166667,  0.98387097,
0.93589744,  0.89915966,  0.88489209,  0.89285714,  0.8989899 ,
0.84415584,  0.02857143,  0.98496241,  0.96590909,  0.91240876,
0.94545455,  0.90526316,  0.67924528,  0.109375  ,  0.89090909,
0.98795181,  0.02777778,  0.94642857,  0.18918919,  0.67605634,
0.55      ,  0.53030303,  0.45098039,  0.05454545,  0.96363636,
0.06122449,  0.98039216,  0.99443414,  0.98688525,  0.27586207,
0.1025641 ,  0.11764706,  0.05128205,  0.81395349,  0.69387755,
0.98611111,  0.99466192,  0.98951782,  0.98723404,  0.97122302,
0.97183099,  0.75862069,  0.98550725,  0.97368421,  0.56      ,
0.98657718,  0.90196078,  0.77419355,  0.65625   ,  0.87012987,
0.25581395,  0.21153846,  0.71794872,  0.52      ,  0.02222222,
0.15625   ,  0.05      ,  0.10714286,  0.8902439 ,  0.79310345,
0.65384615,  0.94174757,  0.65116279,  0.59459459,  0.58823529,
0.0952381 ,  0.10638298,  0.20430108,  0.89361702,  0.65217391,
0.84090909,  0.92753623,  0.89156627,  0.89333333,  0.890625  ,
0.38709677,  0.60869565,  0.65853659,  0.42105263,  0.88405797,
0.92473118,  0.86486486,  0.02985075,  0.40625   ,  0.97916667,
0.52631579,  0.19230769,  0.98571429,  0.53571429,  0.12765957,
0.97333333,  0.67857143,  0.93506494,  0.88976378,  0.87037037,
0.81081081,  0.12244898,  0.51724138,  0.89502762,  0.51851852,
0.93181818,  0.82692308,  0.73529412,  0.22857143,  0.62962963,
0.31034483,  0.9787234 ,  0.96078431,  0.45714286,  0.98033708,
0.93877551,  0.86904762,  0.98268398,  0.98850575,  0.98148148,
0.56756757,  0.99145299,  0.17948718,  0.71641791,  0.91111111,
0.82653061,  0.98734177,  0.984375  ,  0.58064516,  0.47826087,
0.44      ,  0.39130435,  0.20454545,  0.98351648,  0.95714286,
0.96503497,  0.86263736,  0.12      ,  0.76595745,  0.86440678,
0.89873418,  0.91525424,  0.91071429,  0.88636364,  0.48275862,
0.03571429,  0.98591549,  0.69090909,  0.9516129 ,  0.48780488,
0.13793103,  0.08108108,  0.11904762,  0.80597015,  0.9273743 ,
0.89308176,  0.3960396 ,  0.98245614,  0.99415205,  0.98969072,
0.96721311,  0.64788732,  0.23333333,  0.99196787,  0.87603306,
0.86567164,  0.87096774,  0.83269962,  0.84057971,  0.82978723,
0.78333333,  0.80434783,  0.78787879,  0.95121951,  0.13157895,
0.96923077,  0.08571429,  0.98701299,  0.775     ,  0.90163934,
```

```
0.9245283 ,  0.34285714,  0.14814815,  0.83529412,  0.79487179,
0.74666667,  0.73239437,  0.74285714,  0.63855422,  0.63414634,
0.04347826,  0.84782609,  0.81632653,  0.94871795,  0.04301075,
0.22580645,  0.98924731,  0.84375   ,  0.94047619,  0.91891892,
0.85416667,  0.67307692,  0.97468354,  0.74545455,  0.7311828 ,
0.45945946,  0.97938144,  0.92405063,  0.97101449,  0.68493151,
0.58      ,  0.79104478,  0.68888889,  0.99278846,  0.87179487,
0.36111111,  0.36206897,  0.21621622,  0.08510638,  0.62745098,
0.48387097,  0.25806452,  0.98633257,  0.97761194,  0.9789916 ,
0.97530864,  0.96470588,  0.95588235,  0.89705882,  0.85483871,
0.72463768,  0.05405405,  0.8630137 ,  0.16129032,  0.29166667,
0.70454545,  0.23255814,  0.94623656,  0.95412844,  0.75555556,
0.67647059,  0.08      ,  0.98837209,  0.4137931 ,  0.59375   ,
0.52777778,  0.48      ,  0.46551724,  0.34146341,  0.36619718,
0.24137931,  0.30188679,  0.265625  ,  0.09756098,  0.13559322,
0.91440953,  0.91509434,  0.89622642,  0.86086957,  0.85436893,
0.85245902,  0.81609195,  0.8030303 ,  0.78      ,  0.77647059,
0.79545455,  0.76521739,  0.77966102,  0.72321429,  0.72972973,
0.67741935,  0.62222222,  0.98652291,  0.78873239,  0.26086957,
0.71875   ,  0.39285714,  0.87804878,  0.69444444,  0.79411765,
0.992     ,  0.97647059,  0.31578947,  0.31707317,  0.88679245,
0.79591837,  0.9261745 ,  0.8629174 ,  0.98666667,  0.26923077,
0.17857143,  0.38235294,  0.99180328,  0.15942029,  0.90277778,
0.36      ,  0.98507463,  0.7721519 ,  0.04651163,  0.68965517,
0.95890411,  0.06766917,  0.56603774,  0.69767442,  0.93442623,
0.97807757,  0.52173913,  0.75471698,  0.70967742,  0.98076923,
0.23809524,  0.95522388,  0.87142857,  0.74418605,  0.83783784,
0.75510204,  0.59090909,  0.89711934,  0.87301587,  0.89795918,
0.73493976,  0.99122807,  0.96644295,  0.95876289,  0.86046512,
0.07954545,  0.76666667,  0.16216216,  0.02739726,  0.09677419,
0.27659574,  0.83636364,  0.65306122,  0.53521127,  0.97580645,
0.93478261,  0.7755102 ,  0.98672566,  0.99619772,  0.9876161 ,
0.97169811,  0.92957746,  0.97534247,  0.97123894,  0.97191011,
0.97969543,  0.96478873,  0.95491803,  0.03508772,  0.73584906,
0.96341463,  0.69135802,  0.61764706,  0.74358974,  0.92428198,
0.93421053,  0.78723404,  0.37931034,  0.95762712,  0.92783505,
0.16      ,  0.34615385,  0.76388889,  0.63461538,  0.68518519,
0.67567568,  0.675     ,  0.98394495,  0.40540541,  0.57575758,
0.89189189,  0.86330935,  0.18604651,  0.98897059,  0.9673913 ,
0.9379562 ,  0.93243243,  0.98347107,  0.19444444,  0.91139241,
0.84444444,  0.93220339,  0.968     ,  0.53125   ,  0.71698113,
0.80672269,  0.02325581,  0.21875   ,  0.89519651,  0.71604938,
0.2826087 ,  0.07462687,  0.97887324,  0.58974359,  0.33928571,
0.21818182,  0.06779661,  0.28947368,  0.9625    ,  0.95081967,
0.91549296,  0.10344828,  0.99212598,  0.84848485,  0.07317073,
0.97831325,  0.97972973,  0.96511628,  0.9202454 ,  0.90140845,
0.14492754,  0.37837838,  0.46511628,  0.98765432,  0.98697068,
```

```
0.91875   ,  0.84274194,  0.84693878,  0.828125  ,  0.6969697 ,
0.02439024,  0.99090909,  0.94078947,  0.94666667,  0.98979592,
0.81132075,  0.87654321,  0.15277778,  0.68085106,  0.82022472,
0.88321168,  0.96703297,  0.08955224,  0.31818182,  0.96610169,
0.95302013,  0.94375   ,  0.81578947,  0.98319328,  0.98639456,
0.89719626,  0.99107143,  0.64864865,  0.88947368,  0.78688525,
0.825     ,  0.34482759,  0.95098039,  0.04444444,  0.95804196,
0.18461538,  0.97663551,  0.97196262,  0.72641509,  0.3877551 ,
0.85964912,  0.43333333,  0.34920635,  0.09195402,  0.98529412,
0.36666667,  0.9695122 ,  0.83928571,  0.75675676,  0.5862069 ,
0.98780488,  0.48648649,  0.03448276,  0.04285714,  0.96039604,
0.01098901,  0.43902439,  0.31428571,  0.95327103,  0.98695652,
0.94949495,  0.89230769,  0.83673469,  0.75438596,  0.24242424,
0.25961538,  0.75409836,  0.14634146,  0.9627907 ,  0.30232558,
0.97163121,  0.37037037,  0.0212766 ,  0.5952381 ,  0.995     ,
0.99029126,  0.99689441,  0.9691358 ,  0.74      ,  0.34210526,
0.38596491,  0.97857143,  0.90697674,  0.20588235,  0.97807018,
0.97905759,  0.77570093,  0.11666667,  0.85384615,  0.92380952,
0.76829268,  0.91612903,  0.91025641,  0.99019608,  0.58181818,
0.46875   ,  0.73170732,  0.97321429,  0.70731707,  0.59259259,
0.10810811,  0.89830508,  0.44827586,  0.97457627,  0.8961039 ,
0.54285714,  0.07843137,  0.98251748,  0.90234375,  0.02830189,
0.98947368,  0.98684211,  0.17073171,  0.69724771,  0.976     ,
0.98165138,  0.97590361,  0.90438247,  0.40740741,  0.8490566 ,
0.96410256,  0.9380531 ,  0.77358491,  0.17777778,  0.44117647,
0.03076923,  0.87755102,  0.0962963 ,  0.91836735,  0.87951807,
0.80851064,  0.99141104])
```

[6]:
```
data['%Upvote'] = pd.cut(data['Helpful%'], bins = [-1,0,0.2,0.4,0.6,0.8,1],␣
↪labels=['Empty','0-20%', '20-40%', '40-60%', '60-80%', '80-100%'])
data['%Upvote']
```

[6]:
```
0         80-100%
1             NaN
2         80-100%
3         80-100%
4             NaN
           ...
568449        NaN
568450        NaN
568451    80-100%
568452    80-100%
568453        NaN
Name: %Upvote, Length: 568454, dtype: category
Categories (6, object): ['Empty' < '0-20%' < '20-40%' < '40-60%' < '60-80%' <
'80-100%']
```

[7]:
```
data
```

[7]:

	Id	ProductId	UserId	ProfileName
0	1	B001E4KFG0	A3SGXH7AUHU8GW	delmartian
1	2	B00813GRG4	A1D87F6ZCVE5NK	dll pa
2	3	B000LQOCH0	ABXLMWJIXXAIN	Natalia Corres "Natalia Corres"
3	4	B000UA0QIQ	A395BORC6FGVXV	Karl
4	5	B006K2ZZ7K	A1UQRSCLF8GW1T	Michael D. Bigham "M. Wassir"
…	…	…	…	…
568449	568450	B001EO7N10	A28KG5XORO54AY	Lettie D. Carter
568450	568451	B003S1WTCU	A3I8AFVPEE8KI5	R. Sawyer
568451	568452	B004I613EE	A121AA1GQV751Z	pksd "pk_007"
568452	568453	B004I613EE	A3IBEVCTXKNOH	Kathy A. Welch "katwel"
568453	568454	B001LR2CU2	A3LGQPJCZVL9UC	srfell17

	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time
0	1	1	5	1303862400
1	0	0	1	1346976000
2	1	1	4	1219017600
3	3	3	2	1307923200
4	0	0	5	1350777600
…	…	…	…	…
568449	0	0	5	1299628800
568450	0	0	2	1331251200
568451	2	2	5	1329782400
568452	1	1	5	1331596800
568453	0	0	5	1338422400

	Summary
0	Good Quality Dog Food
1	Not as Advertised
2	"Delight" says it all
3	Cough Medicine
4	Great taffy
…	…
568449	Will not do without
568450	disappointed
568451	Perfect for our maltipoo
568452	Favorite Training and reward treat
568453	Great Honey

	Text	Helpful%	%Upvote
0	I have bought several of the Vitality canned d…	1.0	80-100%
1	Product arrived labeled as Jumbo Salted Peanut…	-1.0	NaN
2	This is a confection that has been around a fe…	1.0	80-100%
3	If you are looking for the secret ingredient i…	1.0	80-100%
4	Great taffy at a great price. There was a wid…	-1.0	NaN

```
…                                                      …       …       …
568449  Great for sesame chicken..this is a good if no…    -1.0     NaN
568450  I'm disappointed with the flavor. The chocolat…    -1.0     NaN
568451  These stars are small, so you can give 10-15 o…     1.0  80-100%
568452  These are the BEST treats for training and rew…     1.0  80-100%
568453  I am very satisfied ,product is as advertised,…    -1.0     NaN

[568454 rows x 12 columns]
```

1.3 Q.Analyze upvotes for diff scores

[8]:
```
data.groupby(['Score','%Upvote']).agg('count')
```

[8]:

Score	%Upvote	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Time	Summary	Text	Helpful%
1	Empty	8060	8060	8060	8060	8060	8060	8060	8060	8060	8060
	0-20%	2338	2338	2338	2338	2338	2338	2338	2338	2338	2338
	20-40%	4649	4649	4649	4649	4649	4649	4649	4649	4649	4649
	40-60%	6586	6586	6586	6586	6586	6586	6586	6586	6586	6586
	60-80%	5838	5838	5838	5836	5838	5838	5838	5838	5838	5838
	80-100%	12531	12531	12531	12531	12531	12531	12531	12531	12531	12531
2	Empty	4234	4234	4234	4234	4234	4234	4234	4234	4234	4234
	0-20%	762	762	762	762	762	762	762	762	762	762
	20-40%	1618	1618	1618	1618	1618	1618	1618	1618	1618	1618
	40-60%	3051	3051	3051	3051	3051	3051	3051	3051	3051	3051
	60-80%	2486	2486	2486	2486	2486	2486	2486	2486	2486	2486
	80-100%	7014	7014	7014	7014	7014	7014	7014	7014	7014	7014
3	Empty	5062	5062	5062	5062	5062	5062	5062	5062	5062	5062
	0-20%	474	474	474	474	474	474	474	474	474	474
	20-40%	1506	1506	1506	1506	1506	1506	1506	1506	1506	1506
	40-60%	3384	3384	3384	3384	3384	3384	3384	3384	3384	3384
	60-80%	2754	2754	2754	2754	2754	2754	2754	2754	2754	2754
	80-100%	11037	11037	11037	11037	11037	11037	11037	11037	11037	11037
4	Empty	4780	4780	4780	4780	4780	4780	4780	4780	4780	4780
	0-20%	116	116	116	116	116	116	116	116	116	116
	20-40%	909	909	909	909	909	909	909	909	909	909
	40-60%	3185	3185	3185	3185	3185	3185	3185	3185	3185	3185
	60-80%	2941	2941	2941	2941	2941	2941	2941	2941	2941	2941
	80-100%	26707	26707	26707	26707	26707	26707	26707	26707	26707	26707
5	Empty	11638	11638	11638	11638	11638	11638	11638	11638	11638	11638
	0-20%	432	432	432	432	432	432	432	432	432	432
	20-40%	2275	2275	2275	2275	2275	2275	2275	2275	2275	2275
	40-60%	10312	10312	10312	10312	10312	10312	10312	10312	10312	10312
	60-80%	11060	11060	11060	11060	11060	11060	11060	11060	11060	11060
	80-100%	140661	140661	140661	140659	140661	140661	140661	140661	140661	140661

Score	%Upvote					
1	Empty	8060	8060	8060	8060	8060
	0-20%	2338	2338	2338	2338	2338
	20-40%	4649	4649	4649	4649	4649
	40-60%	6586	6586	6586	6586	6586
	60-80%	5838	5838	5838	5838	5838
	80-100%	12531	12531	12531	12531	12531
2	Empty	4234	4234	4234	4234	4234
	0-20%	762	762	737	762	762
	20-40%	1618	1618	1618	1618	1618
	40-60%	3051	3051	3051	3051	3051
	60-80%	2486	2486	2486	2486	2486
	80-100%	7014	7014	7014	7014	7014
3	Empty	5062	5062	5062	5062	5062
	0-20%	474	474	474	474	474
	20-40%	1506	1506	1506	1506	1506
	40-60%	3384	3384	3384	3384	3384
	60-80%	2754	2754	2754	2754	2754
	80-100%	11037	11037	11036	11037	11037
4	Empty	4780	4780	4780	4780	4780
	0-20%	116	116	116	116	116
	20-40%	909	909	909	909	909
	40-60%	3185	3185	3185	3185	3185
	60-80%	2941	2941	2941	2941	2941
	80-100%	26707	26707	26707	26707	26707
5	Empty	11638	11638	11638	11638	11638
	0-20%	432	432	432	432	432
	20-40%	2275	2275	2275	2275	2275
	40-60%	10312	10312	10312	10312	10312
	60-80%	11060	11060	11060	11060	11060
	80-100%	140661	140661	140661	140661	140661

[9]:
```
data_s = data.groupby(['Score','%Upvote']).agg({'Id':'count'}).reset_index()
data_s
```

[9]:

	Score	%Upvote	Id
0	1	Empty	8060
1	1	0-20%	2338
2	1	20-40%	4649
3	1	40-60%	6586
4	1	60-80%	5838
5	1	80-100%	12531
6	2	Empty	4234
7	2	0-20%	762
8	2	20-40%	1618
9	2	40-60%	3051
10	2	60-80%	2486

```
11      2  80-100%    7014
12      3    Empty    5062
13      3     0-20%     474
14      3    20-40%    1506
15      3    40-60%    3384
16      3    60-80%    2754
17      3  80-100%   11037
18      4    Empty    4780
19      4     0-20%     116
20      4    20-40%     909
21      4    40-60%    3185
22      4    60-80%    2941
23      4  80-100%   26707
24      5    Empty   11638
25      5     0-20%     432
26      5    20-40%    2275
27      5    40-60%   10312
28      5    60-80%   11060
29      5  80-100%  140661
```

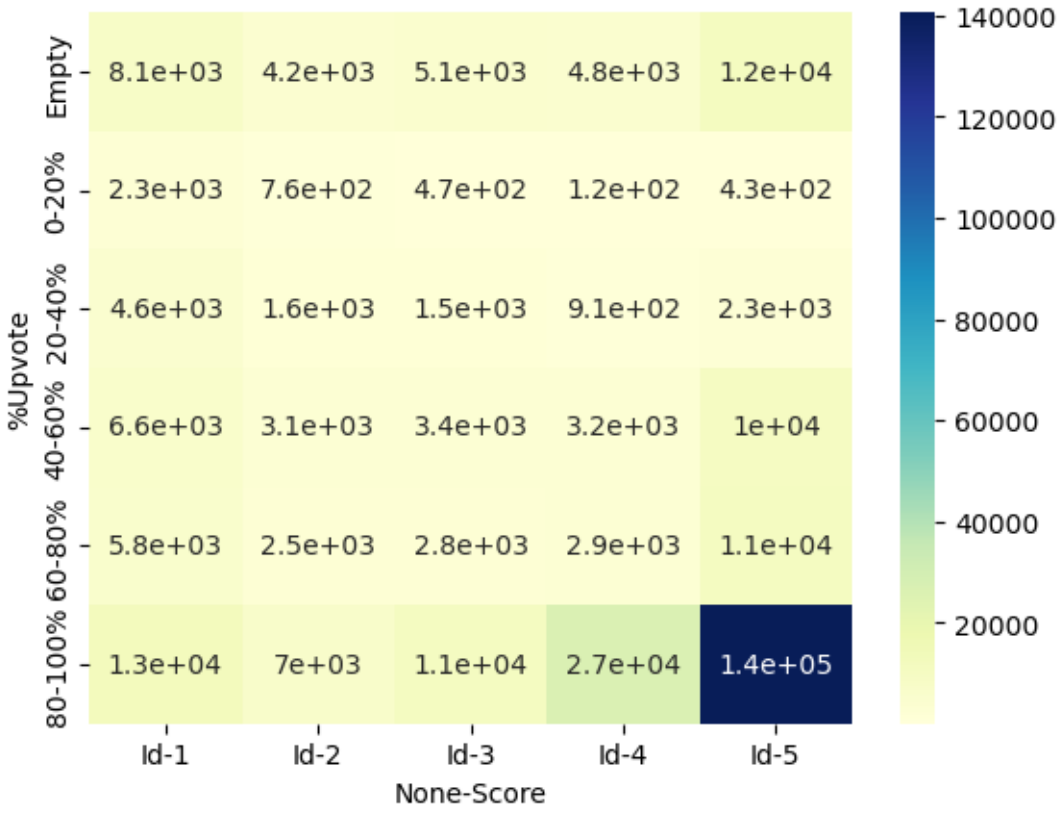
1.4 Q.Create pivot table and heatmap

[10]:
```
pivot = data_s.pivot(index = '%Upvote',columns='Score')
pivot
```

[10]:
```
            Id
Score        1     2      3      4       5
%Upvote
Empty     8060  4234   5062   4780   11638
0-20%     2338   762    474    116     432
20-40%    4649  1618   1506    909    2275
40-60%    6586  3051   3384   3185   10312
60-80%    5838  2486   2754   2941   11060
80-100%  12531  7014  11037  26707  140661
```

[11]:
```
import seaborn as sns
```

[12]:
```
sns.heatmap(pivot, annot = True, cmap='YlGnBu');
```



1.5 Q.Apply Bag of Words on data

[13]:
```
data['Score'].unique()
```

[13]:
```
array([5, 1, 4, 2, 3], dtype=int64)
```

[14]:
```
df1 = data[data['Score'] != 3]
df1
```

[14]:

	Id	ProductId	UserId	ProfileName	\
0	1	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	
1	2	B00813GRG4	A1D87F6ZCVE5NK	dll pa	
2	3	B000LQOCH0	ABXLMWJIXXAIN	Natalia Corres "Natalia Corres"	
3	4	B000UA0QIQ	A395BORC6FGVXV	Karl	
4	5	B006K2ZZ7K	A1UQRSCLF8GW1T	Michael D. Bigham "M. Wassir"	
…	…	…	…	…	
568449	568450	B001EO7N10	A28KG5XORO54AY	Lettie D. Carter	
568450	568451	B003S1WTCU	A3I8AFVPEE8KI5	R. Sawyer	
568451	568452	B004I613EE	A121AA1GQV751Z	pksd "pk_007"	
568452	568453	B004I613EE	A3IBEVCTXKNOH	Kathy A. Welch "katwel"	
568453	568454	B001LR2CU2	A3LGQPJCZVL9UC	srfell17	

```
        HelpfulnessNumerator  HelpfulnessDenominator  Score        Time  \
0                          1                       1      5  1303862400
1                          0                       0      1  1346976000
2                          1                       1      4  1219017600
3                          3                       3      2  1307923200
4                          0                       0      5  1350777600
…                        …                       …    …          …
568449                     0                       0      5  1299628800
568450                     0                       0      2  1331251200
568451                     2                       2      5  1329782400
568452                     1                       1      5  1331596800
568453                     0                       0      5  1338422400

                                 Summary  \
0                    Good Quality Dog Food
1                        Not as Advertised
2                    "Delight" says it all
3                           Cough Medicine
4                              Great taffy
…                                      …
568449                 Will not do without
568450                       disappointed
568451            Perfect for our maltipoo
568452  Favorite Training and reward treat
568453                       Great Honey

                                                     Text  Helpful%  %Upvote
0       I have bought several of the Vitality canned d…       1.0  80-100%
1       Product arrived labeled as Jumbo Salted Peanut…      -1.0      NaN
2       This is a confection that has been around a fe…       1.0  80-100%
3       If you are looking for the secret ingredient i…       1.0  80-100%
4       Great taffy at a great price.  There was a wid…      -1.0      NaN
…                                                     …       …      …
568449  Great for sesame chicken..this is a good if no…      -1.0      NaN
568450  I'm disappointed with the flavor. The chocolat…      -1.0      NaN
568451  These stars are small, so you can give 10-15 o…       1.0  80-100%
568452  These are the BEST treats for training and rew…       1.0  80-100%
568453  I am very satisfied ,product is as advertised,…      -1.0      NaN

[525814 rows x 12 columns]
```

[15]:
```
# score is the dependent variable here
X = df1['Text']
```

[16]:
```
y_dict={1:0,2:0,4:1,5:1}
y = df1['Score'].map(y_dict)
```

[17]:
```
#convert text to vector using NLP
from sklearn.feature_extraction.text import CountVectorizer
```

[18]:
```
#after countvectorizeration the feature no. changes from x to 114969
c = CountVectorizer(stop_words='english')
X_c = c.fit_transform(X)
X_c.shape
```

[18]: (525814, 114969)

1.6 Q.Check Model Accuracy

[19]:
```
from sklearn.model_selection import train_test_split
```

[20]:
```
# default training size = 0.75
X_train, X_test, y_train, y_test = train_test_split(X_c, y)
```

[21]:
```
from sklearn.linear_model import LogisticRegression
```

[22]:
```
log = LogisticRegression(max_iter=1000)
ml = log.fit(X_train,y_train)
```

[23]:
```
ml.score(X_test,y_test)
```

[23]: 0.9398725029287812

1.7 Q.Fetch Top 20 positive words & Top 20 negative words

[24]:
```
w = c.get_feature_names_out()
w
```

[24]: array(['00', '000', '0000', …, 'être', 'île', 'ît'], dtype=object)

[25]:
```
coef = ml.coef_.tolist()[0]
coef
```

[25]:
```
[-0.4045104921437917,
 -0.2349806277079232,
 0.7650130688479836,
 -0.013268855356513653,
 6.256438011803914e-05,
 -0.06998778646891057,
 2.3312339443976707e-05,
 -0.06574080563857287,
 0.0005607250555351886,
 5.170338578455725e-06,
 0.0,
 -9.157431216370397e-05,
```

```
0.0,
0.03962141247706786,
-0.8674153961462652,
0.029104086124453386,
1.071172371345589e-05,
0.00511804928252063,
0.0,
3.5080602057194366e-07,
-0.0007992307425265528,
0.0,
0.022604356514291733,
0.022604356514291733,
0.022604356514291733,
0.022604356514291733,
0.022604356514291733,
0.022604356514291733,
0.022604356514291733,
0.022604356514291733,
0.022604356514291733,
-3.0042887194426984e-06,
0.16958961001612732,
1.5291027632375746e-06,
-0.26996656593461954,
-4.921420867968709e-06,
2.1733642110114007e-06,
0.0035769082735550898,
0.09302804439127715,
3.1242595434541716e-07,
-0.23540705244461674,
0.001248805485944882,
0.00014140470915588801,
6.16873270757308e-05,
0.0,
7.47078139083142e-07,
0.004219023904328949,
2.211220292838409e-06,
1.5291027632375746e-06,
-0.0036908593902379885,
0.004352219157009841,
0.0017979675829074538,
0.0017979675829074538,
0.0,
8.661873649367882e-07,
4.7502733847376296e-08,
1.3339062820930578e-05,
0.0003565513022082112,
-0.711989293424269,
```

```
0.10501897424061578,
-0.336322282796114,
0.10066802478060491,
0.05298215865129302,
0.0,
0.009052459093886163,
0.2655171251523488,
0.0013537929389325915,
0.018448776328339245,
-0.06630153069410229,
0.47524684402680556,
0.13337331151056286,
0.02622443502491088,
-0.000503628548885993,
-0.00033915972488756795,
5.4506416893776e-05,
0.00011642434474348682,
-7.132593078215307e-05,
0.041537142524909955,
0.041537142524909955,
-9.286455657091706e-05,
2.4375265921720014e-05,
2.4375265921720014e-05,
-0.06995935195304781,
0.0,
0.005149055756862622,
0.01019698616167201,
0.37167403630568924,
6.04671737041886e-06,
0.003475394333278474,
-1.2902444072410114e-06,
0.1450773016968839,
0.00014681156320237593,
2.9232678551895882e-05,
0.0,
-0.2472497351540583,
3.866989626342186e-06,
0.0,
5.6393191070390325e-06,
-7.764073170720595e-06,
-0.002105378793438592,
-0.001052689396719296,
-0.001052689396719296,
4.036598729226502e-05,
0.1637916216514349,
0.1076145812689362,
0.01594323128628808,
```

```
0.0,
-0.26996656593461954,
0.0036869874147087175,
0.0,
-0.00018208112359385803,
4.4018135082983405e-06,
-1.1537929101231349e-08,
0.3690314306361781,
0.0,
-1.125262813063894,
-0.05138181623776628,
0.0,
-0.001052689396719296,
-2.6244263615372923e-05,
-0.17406476362348725,
-0.001052689396719296,
0.00012795004252836952,
0.00012906814633736237,
0.023984479875254525,
0.0,
0.00011349597377624158,
0.010412453426216632,
-0.003158068190158739,
0.0,
6.199695450844558e-06,
4.4861393682271635e-07,
4.789736583472552e-06,
0.0,
2.5802243013837134e-06,
5.6393191070390325e-06,
2.1538131891119588e-07,
0.0001211050935222046,
7.47078139083142e-07,
-0.001052689396719296,
0.004150426029486269,
0.0,
-0.3066907546680891,
-0.003158068190158739,
0.0,
-0.001052689396719296,
0.00019232389264140797,
0.13233877030706245,
0.0010256369179347591,
0.032197704787792984,
0.050469062456940876,
0.0013751429829469894,
-9.157431216370397e-05,
```

```
-0.2859556860772267,
4.641470267651736e-06,
-1.0898211622431106e-05,
2.4375265921720014e-05,
0.0008617020980798693,
-9.157431216370397e-05,
3.047755412393689e-05,
4.086589887711135e-05,
0.0,
0.012679200692510561,
1.2465916491983007e-06,
0.07271227981490846,
1.5221750371590615e-07,
2.5585928212742016e-05,
-0.565540145066442,
0.2500313947996132,
-0.002105378793438592,
-0.001052689396719296,
0.0004845602248325423,
-0.001052689396719296,
-0.001052689396719296,
8.909342537474264e-05,
-0.001052689396719296,
-0.002105378793438592,
-0.001052689396719296,
-0.003158068190158739,
-0.002105378793438592,
0.004786097320158187,
1.3021103507749088e-07,
-0.001052689396719296,
0.06251715534398672,
-0.0004899298626995971,
0.0001172480022594230 6,
0.040970836052830084,
0.002208885132409379,
5.98175186236147e-06,
0.040970836052830084,
0.18320398549215264,
-0.38671842082320235,
0.0,
8.229406171086008e-07,
0.005896289375749083,
1.890642783070608e-05,
-0.41473578962768876,
2.2978393157021376e-06,
0.0,
8.229406171086008e-07,
```

```
0.00010158190245340805,
0.22433436909660506,
1.5539490404594322e-05,
0.0004516801385960345,
0.00735854459805898,
0.0018980938150452884,
0.03729186236939083,
-1.2902444072410114e-06,
0.0812887626269445,
0.2154533251702751,
-1.2902444072410114e-06,
0.0023244072779978952,
4.651592037800101e-08,
-0.8425457204039911,
-5.1609776289640454e-06,
-4.504221194147834e-05,
-0.001052689396719296,
0.22314889879110972,
-1.2902444072410114e-06,
-0.001052689396719296,
-0.001052689396719296,
0.1674058262022526,
-0.001052689396719296,
0.0,
-0.001052689396719296,
-0.001052689396719296,
0.0005542738334990256,
0.0001623452069888051,
0.1637916216514349,
-0.001052689396719296,
-0.002105378793438592,
-0.001052689396719296,
-0.001052689396719296,
-0.0002449649313497986,
-0.022552631906205572,
-0.1488824089744135,
7.009705613921153e-06,
0.0,
0.001454069449673594,
0.003941276526604113,
-0.04944290168836252,
0.0,
5.151158840793597e-05,
1.0815957848363595e-08,
-0.03411909685851822,
0.01258612229066007,
-0.003158068190158739,
```

```
-0.001052689396719296,
-0.002105378793438592,
-1.2902444072410114e-06,
-0.1359614012236564,
-0.001052689396719296,
-0.001052689396719296,
-0.001052689396719296,
-0.002105378793438592,
0.0003565513022082112,
-0.002105378793438592,
0.2115319768509137,
-0.436785696839693,
-0.001052689396719296,
-0.002105378793438592,
-0.002105378793438592,
-1.2902444072410114e-06,
0.08964245519405595,
0.0,
9.835787787782796e-07,
-0.14389263310417522,
0.0,
0.0,
0.0,
1.1466739594872665e-05,
1.3629126510852172e-06,
7.04543494803056e-06,
0.09488830150477962,
1.77311215084011e-06,
0.03994850170901904,
0.0,
-0.003943611027945778,
0.02794302502062159,
4.3883896863677845e-06,
-0.10031697212034676,
0.8606755136408487,
0.0,
-0.06717349635180264,
-0.22033372752023428,
0.0,
-0.5586070805651631,
0.0,
0.0,
-0.02170575893789872,
-0.001052689396719296,
0.0023232071769632803,
-0.001052689396719296,
-0.001052689396719296,
```

```
-0.001052689396719296,
5.370791305010571e-07,
2.4250791965390323e-07,
0.004857888172950368,
0.0018980938150452884,
0.03049834511646832,
0.009906268280423465,
5.758616351961845e-07,
0.0,
0.0,
-0.41473578962768876,
-0.26516466301200586,
0.0002972408532288173,
0.01413399653131652,
0.012984384271870148,
0.00029361566718267293,
3.1612331301354605e-06,
7.653286513553033e-06,
0.0,
0.0,
-0.03864934174625687,
0.031955781332646184,
1.5539490404594322e-05,
1.3529586832313816e-06,
6.909062094085687e-06,
2.0061403236969823e-06,
2.1247278996008655e-06,
0.0,
-0.00016117241853487013,
0.013517802362320414,
8.229406171086008e-07,
4.7989591920729056e-05,
0.20566216320352607,
2.9232678551895882e-05,
0.0,
-0.771771434140577,
0.16687477293833336,
0.0,
0.0,
0.00024763694721636185,
0.0005540835109813813,
-0.0031909544860160063,
0.0,
0.0025922358937254296,
4.852469414614307e-09,
1.5539490404594322e-05,
-8.087817716601951e-06,
```

```
0.0021754613250751405,
0.006696932037198505,
0.06666466158051984,
0.0,
0.0,
8.982767238997056e-06,
4.1399392365072884e-06,
5.515024000764467e-07,
0.0,
0.20566216320352607,
0.0003867653400987826,
-0.41473578962768876,
2.4375265921720014e-05,
-1.2902444072410114e-06,
1.5291027632375746e-06,
-0.0017507781631908837,
-0.5156609339888633,
-4.7678729933612204e-05,
-4.7678729933612204e-05,
0.09425397848290992,
0.007099198076129091,
5.814835595756754e-06,
0.0151696022731299,
0.971170676254115,
-0.3128633094550486,
0.0,
0.037179292277155156,
0.1922449256327494,
-0.17203792944916868,
-0.6334190072923341,
-1.2902444072410114e-06,
5.6393191070390325e-06,
-0.007616587797010877,
0.0,
0.0,
0.004350922650150281,
-0.7788500643413164,
-0.020477743921428884,
-0.020477743921428884,
0.014705771388350199,
-0.05744144540380533,
3.611751040918468e-05,
3.611751040918468e-05,
9.573273155244331e-05,
0.013914315543143268,
4.0436526237553595e-07,
-0.34257526813539907,
```

```
0.09425397848290992,
-0.41473578962768876,
0.06328540434827427,
2.0327684303235686e-06,
2.234930316126346e-06,
-0.023034058164365278,
2.546880045893267e-06,
0.0,
2.211220292838409e-06,
-1.6674932969535953e-07,
-0.740313758454501,
0.3251814819218368,
2.981419922574719e-07,
6.16873270757308e-05,
0.05535594533987546,
0.0,
0.01802454626791767,
0.0008411929665519682,
0.11961668598367084,
0.0002638656804055562,
0.0,
0.0,
3.4239699516619585e-05,
1.5564343034323995e-05,
0.0,
0.0,
0.0,
0.09187173939983188,
0.0,
0.05748453930279836,
0.7070320649133459,
0.00010662178908955396,
0.00010662178908955396,
2.126726065011826e-05,
0.0012337528033456468,
0.00419344110230856,
0.2354818772175075,
0.0,
0.02491933939424143,
0.00838688220461712,
-0.25668023834017256,
0.00419344110230856,
0.002637935995107708,
0.049601789666429745,
-0.48833124707075587,
0.0,
0.029052823066934828,
```

```
-0.08208869827494852,
0.004646290293072098,
0.07965172197277269,
0.0004325475899562083,
0.10003821165253902,
-0.15121005502343568,
-0.9320506059051299,
-0.22682563116324467,
3.246110641184972e-05,
0.983070845038915,
-0.19782118478986316,
-0.010717049220516865,
0.02448884814063333,
0.6442654890625295,
-0.06402480307732981,
0.012345482083549889,
0.03422337091855797,
-1.0018394891596895e-05,
0.0531579517899413,
0.00578440098384,
4.2668621152525364e-07,
0.0,
0.1744230748904043,
0.014154150242320111,
0.0002738030388649902,
0.042806774023841705,
-0.06628322612392071,
-0.010289346234579192,
0.11713624945851049,
0.004172146317161431,
0.09184487891880415,
0.030116095281743,
-3.750511560407491e-05,
-0.22385054921050854,
4.454671268737132e-05,
0.07524469537641464,
0.025861644986125025,
0.04949357917207811,
0.09087887374684642,
0.31674792405696106,
0.0,
0.6464857140649347,
4.627664175831292e-05,
-0.4938815415130637,
0.0028537260643147935,
-0.015248257592572095,
0.0,
```

```
0.001838847313336972,
-0.6098396392515604,
-3.626077675853901e-05,
0.08078237484849851,
-1.2902444072410114e-06,
-0.9127483806520117,
0.20581128920040231,
3.726966744404871e-06,
-1.3655428325525245,
0.0,
0.3432952228354513,
-0.7330063776290788,
0.12832671195127077,
1.8280221998074157e-05,
0.002892313954468794,
0.033817280793504864,
0.033817280793504864,
0.0,
0.0006791388495783762,
-0.2580325365768105,
-1.042270012532999,
0.00019310606090810724,
-0.3274553076140811,
0.0,
-0.0709579463670053,
0.03185393581697698,
0.01807208786466639,
-0.04895099798202998,
0.0,
0.0,
-0.5464840417758288,
-0.06633500316393982,
-0.007368825777035597,
-0.05988638617933406,
0.04985887407026724,
0.00020043348604771097,
3.23645348309521e-05,
0.0007141716247136034,
0.0,
-0.09150916978721789,
-0.002105378793438592,
0.0,
-1.1616507249461565,
-0.24515414821170942,
0.0066557040906870075,
0.008162639044099942,
0.0,
```

```
3.6709852645402016e-06,
0.0,
-0.001052689396719296,
0.0,
0.20275183951318923,
0.030499847886365825,
6.858960137648304e-06,
-0.005482298567373087,
1.57691460334203e-05,
-0.022961987688657797,
1.6096230340963164e-05,
-0.000917657316704878,
0.0,
-0.9939050397281013,
0.0023189521959903394,
-0.14757148116889252,
0.0,
0.0023189521959903394,
0.003791120521049617,
-0.2707422276900112,
0.3583998800278027,
6.378933274527282e-05,
-0.23918421154307365,
0.06720887874597788,
0.3912707495514061,
-0.0009680650726533754,
0.28376451581793294,
0.03478428293985744,
0.009275808783961358,
0.004637904391980679,
0.006956856587971268,
0.027827426351885072,
0.002352209506643694,
0.009275808783961358,
-0.632210871445198,
0.14569310477449537,
1.4063339192591699e-05,
0.027888442537218155,
0.013946970486595961,
0.0023189521959903394,
0.011594760979951978,
0.023189521959903956,
0.006956856587971268,
0.009275808783961358,
0.013913713175942536,
2.7623158447830167e-07,
0.002212371361390975,
```

```
-0.28516484932172514,
0.03014637854787607,
0.013913713175942536,
0.0023189521959903394,
0.0023189521959903394,
0.006956856587971268,
-2.3927391221897193e-05,
0.006956856587971268,
0.0023189521959903394,
-0.010530297246270522,
0.11229438003779342,
0.13887680466207217,
0.0107456854766673,
0.12901104180496056,
-0.31777191447903186,
0.0006117242220914607,
0.011836044608220787,
-0.36244894315508647,
0.00032712662264402045,
0.099464714542031,
-0.39352344283452695,
0.012189443657843845,
0.0027917435184438557,
0.02349828808960462,
0.0016542069744944921,
0.06693473399501466,
0.026258965701832986,
0.0,
0.0017533007071684585,
0.002634911593808459,
-0.00012770464499478593,
0.41394297987068057,
0.002830051908764604,
-0.6764925494844319,
0.30691844538655816,
2.286445651298652e-07,
1.1857246337803412e-05,
0.0,
0.1019469518811948,
4.454671268737132e-05,
0.0,
0.17752524137246364,
-0.9659785370821395,
0.003843418256137654,
2.80826645838156e-07,
0.06437126852937901,
0.04459337411713808,
```

```
0.12268373759159217,
0.0044679768624918915,
0.3906419137108323,
0.00019972847801423544,
0.031329115293368,
0.03885728313474641,
-0.18551165871623343,
6.954461944190792e-05,
-1.5187407127270367,
1.1449546423337814e-06,
0.01892479832251748,
0.03557833097987246,
0.0001405989698636828,
0.0006910193292306613,
-0.022115294553198935,
0.0718608347372169,
0.0,
0.0005997201244720278,
2.7848310602360605e-06,
1.530552610667832,
0.0037584497894217416,
0.06469785505273065,
-0.00013850938366245148,
-0.05303535775455877,
5.754796965075655e-06,
0.009289516396574202,
0.02621816953720715,
0.7786276366604665,
0.23175146182733178,
1.3939864359878886e-07,
-0.03902445097279002,
0.591840486040127,
0.27197242711837566,
0.0009862068311099147,
0.0,
0.0,
0.0023189521959903394,
0.004637904391980679,
-0.22215472048067786,
-8.808969203556629e-05,
-4.4044846017783146e-05,
0.018222004939380067,
5.6393191070390325e-06,
-0.001052689396719296,
-0.0020997394743320526,
1.7317553447263932e-07,
0.023879759057008634,
```

```
0.0,
0.014138872929789489,
0.2033022659994245,
0.3263107852030453,
0.0,
0.0001897524301054637,
0.0001566169503863259,
5.6393191070390325e-06,
0.2115712860540538,
-0.433201601238329,
-0.45425540484684024,
0.03414342496280124,
-0.001052689396719296,
-1.4546836958546734,
2.435991568076571e-05,
-0.30041869770091645,
0.0,
0.00026179714388841656,
0.0,
-0.0017507781631908837,
0.9473699632776115,
-0.001052689396719296,
0.05185109006171764,
0.6099423156909727,
-0.3223568106852106,
0.01540244870426112,
-0.4243506330945111,
0.08451703127592794,
-0.5282702449779625,
0.0,
0.0,
0.0020005349861948085,
0.3803200295838887,
0.3903149646996796,
6.423490455902551e-07,
0.04555764392843036,
7.106665099449653e-07,
0.043852450556663584,
-0.4871757462971334,
0.09985124037420429,
0.4617411172953452,
5.3773424468004875e-05,
0.0006200961105909953,
-0.07416720114711593,
-0.04944290168836252,
0.048281811339422456,
2.5399640108655e-06,
```

```
0.05684707741081073,
0.025861644986125025,
0.0,
-0.20725569995820417,
0.0,
-2.3927391221897193e-05,
4.183496978532394e-05,
-0.34653396239344664,
4.348752635834031e-07,
0.001986176782332819,
-0.46709970305331944,
0.09897403719696264,
-0.6003131206263198,
0.2917084452439449,
-0.4273183020796215,
0.0,
3.439965282916036e-05,
0.04017297318887886,
0.27425623614212613,
7.166934559498918e-06,
0.00011203106976641568,
0.0,
0.09035243957345933,
0.00012771405664145584,
0.15006860796908297,
0.0,
0.22827139484526757,
0.2522893715776632,
0.0001432978231395 1992,
-0.2188345458756207,
0.05307963986487095,
-0.5686197300514945,
-0.013504189289123204,
0.004021199813165971,
-0.14037067780726253,
0.08967476015006742,
0.01568612460266469,
0.0,
0.49167753798627284,
0.08261490875910593,
0.0,
-0.1510873568232382,
-0.08584806264692336,
0.38234716848903294,
0.0,
-7.1785978822756846e-06,
-0.14950038925781958,
```

-0.03013862905417122,
2.595500421726352e-07,
-0.18014813248403821,
-0.001052689396719296,
-0.001052689396719296,
-1.3669573889172e-05,
-0.4994611731067724,
-0.040136153084044106,
-0.0017507781631908837,
0.13368966629058057,
0.0,
0.6530339438604221,
0.019686466664626735,
0.0009477314604919749,
0.04370120252498679,
0.17999937722841292,
0.00629733820983143,
0.08622678257362089,
0.018355778170643797,
1.6757206831978004,
2.965444415878681e-06,
0.22370187249149237,
0.0,
-0.42194326097525353,
0.06564080273515087,
-0.529694306754392,
-0.005482298567373087,
0.1054849323542664,
0.22633528941374356,
0.6186699867905935,
9.298044468708823e-05,
0.05570663421582265,
0.37328027297767197,
0.0,
0.0,
0.0008154247664596244,
9.298044468708823e-05,
-0.6251891021054388,
0.0,
9.298044468708823e-05,
0.0,
9.598845774202134e-07,
9.598845774202134e-07,
0.1683860437802414,
-0.20065612808625496,
-0.004563521019935787,
0.02818681854871744,

0.04701605500534266,
0.0,
0.3847389441294841,
0.0,
1.323406532973676e-07,
-0.19551167247455128,
0.0025860251138373307,
4.7502733847376296e-08,
0.002884019415411574,
-0.5288092094865497,
0.0004503076897673921,
0.18204758585698894,
0.0031186085407276577,
0.7603723877852802,
0.03851542964740828,
0.045730331064782415,
-0.0023714226061834257,
0.0,
0.0023138717011824375,
0.005840753401171951,
0.00015532013278492446,
3.7637584705732256e-05,
0.04474167115810474,
-0.20522966941592108,
0.1180245193023001,
-0.14966493191585636,
0.11426185904908435,
0.13546491118448428,
0.0,
0.5163361306321945,
8.380551913218187e-07,
-0.41060256047006033,
-0.1976098523930035,
0.00019047760069097454,
-0.1976098523930035,
-0.02213581990750665,
0.0,
0.03138563794935831,
-0.14966493191585636,
0.0,
0.008792091636035757,
-1.6674932969535953e-07,
2.476130557210853e-05,
0.04719589881464389,
-0.6116496151558097,
0.018485486951604128,
-0.04156133160535109,

```
0.10963283509081495,
0.014432968305730725,
0.13955759207806429,
-0.0830784107149972,
0.022326515249419553,
0.00036479866404103597,
0.019686466664626735,
0.0002225490417627515,
0.005593426342053205,
0.03437872984221831,
0.015805955690386578,
0.06693473399501466,
0.17778797425844792,
0.6278728833833673,
0.0,
0.0,
-0.6334190072923341,
0.0,
0.002718172366637891,
-0.23247925083555673,
0.07286482433012387,
0.012527826718262811,
0.0,
0.00013898991365204753,
0.0005389447249604381,
0.017829128047615633,
-0.1790127218712136,
0.11631330484175917,
0.3819598891836329,
-0.43142137749690307,
0.10824355045008491,
-0.6477154486467708,
-0.17758931087017166,
0.02956203941893672,
0.3341978060577247,
0.0027347355950456684,
0.24252534235461906,
0.28943559891830095,
5.783597779767837e-07,
-0.6591093624548717,
0.0017187670846008933,
-0.37799324642938903,
0.0003623530653439882,
0.8018738135714505,
0.10481769953121188,
0.0,
-0.1251154120053335,
```

```
-0.21456517031153977,
0.01311742694690808,
0.0,
0.003228299996674853,
-0.22430280765589128,
0.7054979140551628,
-4.360809172277683e-05,
0.273728280326894,
2.143105825560985e-06,
0.000934560235679036,
0.0036078491698821116,
0.0,
-0.01748262059991479,
0.15198488395392237,
0.12416657020203153,
-0.05303535775455877,
0.0802996169570379,
-0.4377078239844874,
0.056835060554208026,
0.03473755122598096,
0.008218284759175772,
-0.30531620739145865,
0.5893101752790394,
-0.5009024474158343,
0.16501779850679515,
0.6914809288687404,
0.3754653260286691,
0.00021026877510729513,
0.0,
0.0017464374792619914,
0.005636625234232588,
-0.08402694357046317,
-0.002682365922420337,
-0.002682365922420337,
-1.4340190312497414,
0.00010662178908955396,
0.0,
0.0007862238454579564,
0.0,
0.42382833296540157,
-0.2791701939587124,
0.0,
0.3333463547829276,
0.005503304519388795,
0.4416175660143125,
-0.29538486050706947,
0.010808540877042402,
```

```
-0.23663336648222352,
0.0,
-0.09299302493296202,
0.010572830966022954,
0.00011345564524768778,
0.00022691129049537555,
0.00011345564524768778,
0.33496610325620907,
0.00025926633892340255,
5.227609713817068e-07,
5.772984877644688e-05,
0.0767342074568431,
0.27934635880040815,
-0.17295790624531635,
0.05185109006171764,
3.5558948977317274e-05,
-0.18450548439982659,
-0.18450548439982659,
0.0,
0.25940732930126015,
3.277224871578536e-05,
0.00024294957336791408,
0.03210231669654468,
-0.08229852430963912,
-0.0954126515006081,
0.0008930813039107937,
1.3671240692744856e-05,
2.5735961289061037e-05,
-0.005924432821403313,
-0.0034122035279263094,
-6.669295092188283e-06,
0.1393756007430551,
3.3643912233277453e-07,
-0.86013182012089,
4.317086798120845e-07,
0.0,
0.0,
-0.00563831899219322,
0.00017090015937480926,
-0.009102444385841409,
0.0,
0.0,
0.0,
0.000566957641696707,
0.0,
-1.1061542506392796,
0.00025696216850736844,
```

```
 0.00029891251356740494,
 …]
```

[26]:
```
coef_data = pd.DataFrame({'Word':w, 'Coefficient':coef})
coef_data
```

[26]:
```
           Word  Coefficient
0            00    -0.404510
1           000    -0.234981
2          0000     0.765013
3        000001    -0.013269
4         00001     0.000063
…          …           …
114964   çaykur     0.002136
114965    çelem    -0.162245
114966    être     0.007231
114967     île     0.000596
114968      ît     0.000830

[114969 rows x 2 columns]
```

[27]:
```
coef_data = coef_data.sort_values(['Coefficient', 'Word'], ascending=False)
coef_data
```

[27]:
```
                 Word  Coefficient
27198         chedder     3.421853
41175   emeraldforest     3.351335
96145         solving     3.104422
80600      pleasantly     3.047237
20268          bertie     2.938561
…                …           …
76597       overrated    -2.880812
113164          worst    -2.939393
106852   unacceptable    -2.960678
94813      skyrocketd    -2.988298
76621      oversalted    -3.312547

[114969 rows x 2 columns]
```

[28]:
```
coef_data.head(20)
```

[28]:
```
                 Word  Coefficient
27198         chedder     3.421853
41175   emeraldforest     3.351335
96145         solving     3.104422
80600      pleasantly     3.047237
20268          bertie     2.938561
```

```
114056    yirgacheffe     2.569059
108387     unwrapping     2.522339
21553         blowout     2.507810
79122        perruche     2.491426
28998         cleanup     2.485841
94680          skewed     2.464618
64829         looming     2.436514
55029          hooked     2.393526
105598        tribute     2.319534
57550           infer     2.312419
22257           botch     2.302961
73164        noisette     2.266739
107829      unmatched     2.229105
5865        addicting     2.228704
94667       skeptical     2.224480
```

[29]:
```
coef_data.tail(20)
```

[29]:
```
                Word  Coefficient
7321          allegro    -2.526720
103002        textual    -2.533767
41118     embarrassed    -2.636030
34989       deceptive    -2.644098
13533      b000sqn3og    -2.662965
56229             ick    -2.672414
88351      returnable    -2.684963
54767           holle    -2.721002
86421       redeeming    -2.754980
111257        weakest    -2.755373
37560   disappointing    -2.799255
107383    undrinkable    -2.801467
2570          3095826    -2.807892
38117   dissapointing    -2.823191
23580           budda    -2.855555
76597       overrated    -2.880812
113164          worst    -2.939393
106852   unacceptable    -2.960678
94813      skyrocketd    -2.988298
76621      oversalted    -3.312547
```

1.8 Q.Automate the previous 3 tasks

[30]:
```
def text_fit(X,y,nlp_model,ml_model, coef_show=1):
    X_c = nlp_model.fit_transform(X)
    print('features:{}'.format(X_c.shape[1]))

    X_train, X_test, y_train, y_test = train_test_split(X_c, y)
```

```
    ml=ml_model.fit(X_train,y_train)
    acc = ml.score(X_test,y_test)
    print(acc)

    if coef_show == 1:
        w = c.get_feature_names_out()
        coef = ml.coef_.tolist()[0]
        coef_data = pd.DataFrame({'Word':w, 'Coefficient':coef})
        coef_data = coef_data.sort_values(['Coefficient', 'Word'],␣
↪ascending=False)
        print('\n')
        print('--Top 20 Positive Words--')
        print(coef_data.head(20))
        print('\n')
        print('--Top 20 Negative Words--')
        print(coef_data.tail(20))
```

[31]:
```
text_fit(X,y,c,log)
```

```
features:114969
0.9379250536309279


--Top 20 Positive Words--
                Word  Coefficient
41175   emeraldforest     3.766309
27198         chedder     3.424512
80600      pleasantly     3.012987
94680          skewed     2.812335
96145         solving     2.805302
20268          bertie     2.761146
79122        perruche     2.526277
53585           hears     2.489164
57223        incurred     2.464219
105598        tribute     2.458522
55029          hooked     2.441171
93489      shippments     2.425198
114056    yirgacheffe     2.389492
113138        worries     2.370780
39072        downside     2.346560
94667       skeptical     2.345852
5865         addicting     2.320987
32209      correction     2.301746
75638           oranic     2.300034
56956   inappropriate     2.299635


--Top 20 Negative Words--
```

```
                Word  Coefficient
80711           plot    -2.585094
94813     skyrocketd    -2.590600
58245   insufficient    -2.590864
72573         neuman    -2.607501
23580          budda    -2.650093
103002       textual    -2.692833
56229            ick    -2.746001
34989      deceptive    -2.790583
107494    unfinished    -2.796195
60312       juiciest    -2.807061
37560  disappointing    -2.850858
107383   undrinkable    -2.909671
86421      redeeming    -2.916152
106852  unacceptable    -2.933580
111257       weakest    -2.969648
113164         worst    -2.995628
88351     returnable    -3.049977
76621     oversalted    -3.076070
2318           280mg    -3.210278
2570         3095826    -3.245125
```

1.9 Q.Automate the Predictions

[32]:
```
from sklearn.metrics import confusion_matrix, accuracy_score
```

[33]:
```
def predict(X,y,nlp_model,ml_model):
    X_c = nlp_model.fit_transform(X)
    X_train, X_test, y_train, y_test = train_test_split(X_c, y)
    ml=ml_model.fit(X_train,y_train)
    predictions = ml.predict(X_test)
    cm = confusion_matrix(predictions, y_test)
    print(cm)
    acc = accuracy_score(predictions, y_test)
    print(acc)
```

[34]:
```
c = CountVectorizer()
```

[35]:
```
predict(X,y,c,log)
```

```
[[ 15817   2496]
 [  4668 108473]]
0.9455018485553882
```

```
C:\Users\Abdul Mateen\anaconda3\lib\site-
packages\sklearn\linear_model\_logistic.py:458: ConvergenceWarning: lbfgs failed
to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.
```

```
Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(
```

1.10 Q.Apply more NLP & ML on data

[36]:
```
from sklearn.dummy import DummyClassifier
```

[37]:
```
#count vetorization(no parameter) and dummy Classifier
text_fit(X,y,c,DummyClassifier(),0)
```

```
features:115282
0.8439149816665906
```

[38]:
```
from sklearn.feature_extraction.text import TfidfVectorizer
```

[39]:
```
tfidf = TfidfVectorizer(stop_words='english')
text_fit(X,y,tfidf,log,0)
```

```
features:114969
0.9353690264274955
```

[40]:
```
#tf-idf and Logistic regression
predict(X,y,tfidf,log)
```

```
[[ 14256   2241]
 [  6264 108693]]
0.9353005614131179
```

1.11 Q.Data Preparation for predicting the Upvotes

[41]:
```
df2 = data[data['Score']==5]
df2
```

[41]:

	Id	ProductId	UserId	ProfileName
0	1	B001E4KFG0	A3SGXH7AUHU8GW	delmartian
4	5	B006K2ZZ7K	A1UQRSCLF8GW1T	Michael D. Bigham "M. Wassir"
6	7	B006K2ZZ7K	A1SP2KVKFXXRU1	David C. Sullivan
7	8	B006K2ZZ7K	A3JRGQVEQN31IQ	Pamela G. Williams
8	9	B000E7L2R4	A1MZYO9TZK0BBI	R. James
…	…	…	…	…
568448	568449	B001EO7N10	A1F6BHEYB7R6R7	James Braley
568449	568450	B001EO7N10	A28KG5XORO54AY	Lettie D. Carter
568451	568452	B004I613EE	A121AA1GQV751Z	pksd "pk_007"
568452	568453	B004I613EE	A3IBEVCTXKNOH	Kathy A. Welch "katwel"
568453	568454	B001LR2CU2	A3LGQPJCZVL9UC	srfell17

```
        HelpfulnessNumerator  HelpfulnessDenominator  Score        Time  \
0                          1                       1      5  1303862400
4                          0                       0      5  1350777600
6                          0                       0      5  1340150400
7                          0                       0      5  1336003200
8                          1                       1      5  1322006400
…                        …                       …    …         …
568448                     0                       0      5  1308096000
568449                     0                       0      5  1299628800
568451                     2                       2      5  1329782400
568452                     1                       1      5  1331596800
568453                     0                       0      5  1338422400

                                          Summary  \
0                             Good Quality Dog Food
4                                     Great taffy
6       Great!  Just as good as the expensive brands!
7                           Wonderful, tasty taffy
8                                    Yay Barley
…                                               …
568448                 Very large ground spice jars.
568449                          Will not do without
568451                     Perfect for our maltipoo
568452            Favorite Training and reward treat
568453                                    Great Honey

                                                     Text  Helpful%  %Upvote
0       I have bought several of the Vitality canned d…       1.0  80-100%
4       Great taffy at a great price.  There was a wid…      -1.0      NaN
6       This saltwater taffy had great flavors and was…      -1.0      NaN
7       This taffy is so good.  It is very soft and ch…      -1.0      NaN
8       Right now I'm mostly just sprouting this so my…       1.0  80-100%
…                                                     …       …       …
568448  My only complaint is that there's so much of i…      -1.0      NaN
568449  Great for sesame chicken..this is a good if no…      -1.0      NaN
568451  These stars are small, so you can give 10-15 o…       1.0  80-100%
568452  These are the BEST treats for training and rew…       1.0  80-100%
568453  I am very satisfied ,product is as advertised,…      -1.0      NaN

[363122 rows x 12 columns]
```

[42]:
```
data2 = df2[df2['%Upvote'].isin(['80-100%', '60-80%', '40-60%', '20-40%'])]
data2
```

[42]:
```
            Id   ProductId          UserId                ProfileName  \
0            1  B001E4KFG0  A3SGXH7AUHU8GW                 delmartian
```

```
8            9  B000E7L2R4  A1MZYO9TZK0BBI                    R. James
10          11  B0001PB9FE  A3HDKO7OW0QNK4                Canadian Fan
11          12  B0009XLVG0  A2725IB4YY9JEB      A Poeng "SparkyGoHome"
14          15  B001GVISJM  A2MUGFV2TDQ47K         Lynrie "Oh HELL no"
…          …         …            …                          …
568440  568441  B005ZC0RRO  A2TO5R8QLIITEF                        SAK
568444  568445  B001EO7N10  A2SD7TY3IOX69B  BayBay "BayBay Knows Best"
568445  568446  B001EO7N10  A2E5C8TTAED4CQ               S. Linkletter
568451  568452  B004I613EE  A121AA1GQV751Z               pksd "pk_007"
568452  568453  B004I613EE   A3IBEVCTXKNOH     Kathy A. Welch "katwel"

        HelpfulnessNumerator  HelpfulnessDenominator  Score        Time  \
0                          1                       1      5  1303862400
8                          1                       1      5  1322006400
10                         1                       1      5  1107820800
11                         4                       4      5  1282867200
14                         4                       5      5  1268352000
…                        …                       …    …          …
568440                     1                       1      5  1323734400
568444                     3                       3      5  1245369600
568445                     2                       2      5  1268006400
568451                     2                       2      5  1329782400
568452                     1                       1      5  1331596800

                                              Summary  \
0                                  Good Quality Dog Food
8                                          Yay Barley
10                        The Best Hot Sauce in the World
11      My cats LOVE this "diet" food better than thei…
14                           Strawberry Twizzlers - Yummy
…                                                    …
568440   Delicious, all natural and allergy free treats!
568444                     Best Value for Chinese 5 Spice
568445                                 Five Spice Powder
568451                            Perfect for our maltipoo
568452                Favorite Training and reward treat

                                                     Text  Helpful%  %Upvote
0       I have bought several of the Vitality canned d…       1.0  80-100%
8       Right now I'm mostly just sprouting this so my…       1.0  80-100%
10      I don't know if it's the cactus or the tequila…       1.0  80-100%
11      One of my boys needed to lose some weight and …       1.0  80-100%
14      The Strawberry Twizzlers are my guilty pleasur…       0.8   60-80%
…                                                    …       …       …
568440  Indie Candy's gummies are absolutely delicious…       1.0  80-100%
568444  As a foodie, I use a lot of Chinese 5 Spice po…       1.0  80-100%
568445  You can make this mix yourself, but the Star A…       1.0  80-100%
```

```
568451  These stars are small, so you can give 10-15 o…       1.0  80-100%
568452  These are the BEST treats for training and rew…       1.0  80-100%

[164308 rows x 12 columns]
```

[43]:
```
X = data2['Text']
X
```

[43]:
```
0         I have bought several of the Vitality canned d…
8         Right now I'm mostly just sprouting this so my…
10        I don't know if it's the cactus or the tequila…
11        One of my boys needed to lose some weight and …
14        The Strawberry Twizzlers are my guilty pleasur…
                              …
568440    Indie Candy's gummies are absolutely delicious…
568444    As a foodie, I use a lot of Chinese 5 Spice po…
568445    You can make this mix yourself, but the Star A…
568451    These stars are small, so you can give 10-15 o…
568452    These are the BEST treats for training and rew…
Name: Text, Length: 164308, dtype: object
```

[44]:
```
y_dict = {'80-100%':1, '60-80%':1, '20-40%':0, '40-60%':0}
y = data2['%Upvote'].map(y_dict)
y
```

[44]:
```
0         1.0
8         1.0
10        1.0
11        1.0
14        1.0
          …
568440    1.0
568444    1.0
568445    1.0
568451    1.0
568452    1.0
Name: %Upvote, Length: 164308, dtype: float64
```

[45]:
```
y.value_counts()
```

[45]:
```
1.0    151721
0.0     12587
Name: %Upvote, dtype: int64
```

1.12 Q.Apply Tf-Idf on data

[46]:
```
tfidf = TfidfVectorizer(stop_words='english')
X_c = tfidf.fit_transform(X)
X_c.shape
```

[46]:
```
(164308, 69428)
```

[47]:
```
y.value_counts()
```

[47]:
```
1.0    151721
0.0     12587
Name: %Upvote, dtype: int64
```

1.13 Q.Handle Imbalace data if data is Imbalance

[48]:
```
# requires Tensorflow
#pip install tensorflow
```

[49]:
```
from imblearn.over_sampling import RandomOverSampler
```

[50]:
```
os = RandomOverSampler()
```

[51]:
```
X_train_res, y_train_res = os.fit_resample(X_c,y)
```

[52]:
```
from collections import Counter
```

[53]:
```
print('Original Dataset shape {}'.format(Counter(y)))
print('Resampled Dataset shape {}'.format(Counter(y_train_res)))
```

```
Original Dataset shape Counter({1.0: 151721, 0.0: 12587})
Resampled Dataset shape Counter({1.0: 151721, 0.0: 151721})
```

1.14 Q.Do Cross validation using GridSearchCV & then do predictions

[54]:
```
from sklearn.model_selection import GridSearchCV
```

[55]:
```
q = np.arange(-2,3)
q
```

[55]:
```
array([-2, -1,  0,  1,  2])
```

[56]:
```
grid = {'C' : 10.0 **q,'penalty':['l2']}
```

[57]:
```
# n_jobs = -1 use all CPU resources
clf = GridSearchCV(estimator=log, param_grid=grid, cv = 5, n_jobs=-1,␣
 ↪scoring='f1_macro')
```

[58]:
```
clf.fit(X_train_res, y_train_res)
```

```
C:\Users\Abdul Mateen\anaconda3\lib\site-
packages\sklearn\linear_model\_logistic.py:458: ConvergenceWarning: lbfgs failed
to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(
```

[58]:
```
GridSearchCV(cv=5, estimator=LogisticRegression(max_iter=1000), n_jobs=-1,
             param_grid={'C': array([1.e-02, 1.e-01, 1.e+00, 1.e+01, 1.e+02]),
                         'penalty': ['l2']},
             scoring='f1_macro')
```

[59]:
```
X_train, X_test, y_train, y_test = train_test_split(X_c,y) # since the number␣
↪of features have changed, split again
```

[60]:
```
pred = clf.predict(X_test)
pred
```

[60]:
```
array([0., 1., 1., …, 1., 1., 0.])
```

1.15 Q.Checking Accuracy of cross validated model

[61]:
```
from sklearn.metrics import confusion_matrix
```

[62]:
```
confusion_matrix(y_test, pred)
```

[62]:
```
array([[ 2936,   163],
       [ 3254, 34724]], dtype=int64)
```

[63]:
```
accuracy_score(y_test, pred)
```

[63]:
```
0.9168147625191713
```