Hee Ji Park
CSCI544 (NLP)
09.06.2021

Report - HW1 (CSCI544)

1. Introduction

   1.1 Our goal

   : Our goal is to train sentiment analysis classifiers that can predict the sentiment (positive/negative) for a given review.

   1.2 Datasets

   - This dataset contains real reviews for kitchen products sold on Amazon.
   - Data columns: Index(['marketplace', 'customer_id', 'review_id', 'product_id',

     'product_parent', 'product_title', 'product_category', 'star_rating',

     'helpful_votes', 'total_votes', 'vine', 'verified_purchase',

     'review_headline', 'review_body', 'review_date'], dtype='object')
   - However, we only need to 'review_body' and 'star_rating' columns in order to do sentiment analysis.
   - We create binary labels using the ratings. If the star rating score is 4 or 5, it is assumed to be a positive sentiment (mapped to 1), and if the star rating is rating less than 1 or 2, it is assumed to be a negative sentiment (mapped to 0). Rating 3 is considered a neutral opinion. (But we will discard reviews with the rating 3.)

   1.3 Packages and libraries

   - Pandas
   - Numpy
   - NLTK (Stopword / WordNetLemmatizer)
   - Re
   - BeautifulSoup
   - contractions
   - gzip
   - shutil
   - sklearn (train_test_split / TfidfVectorizer / LinearSVC / MultinomialNB / LogisticRegression / Perceptron / precision_score / recall_score / accuracy_score / f1_score)

2. Three sample reviews in report along with corresponding ratings

Rating	Sample Reviews
1	I bought this blender with high hopes, seeing all those informercials claiming that it blends so smoothly. Well when I purchased it I was hoping to get a well blended fruit and vegetable drink, instead I got a chewy mess, and I feel taken for all the money that I paid for it. The 72 oz pitcher is a big waste of money. I can only use the single serve for a better blend. I wish I

	knew this before I purchased it, I could have paid way less for just a single serve, I will not be fooled again, and I do not recommend to anybody!!! Nothing against amazon, just this blender.
	What should I try to grill by using a palm sized set? Probably the worst product that I have bought from Amazon.
	Both glasses were broken in the box - - very thin glass and not packed well. I was dissappointed. I like the idea and also ordered teh frosted style, they did arrive unbroken but as I said, VERY THIN GLASS -- I don't know that they keep beer cold longer.
2	Hi, I received my new coffeemaker, but the carafe lid was broked... How can I hve a new one on well conditions? I need to order a new one? Thanks for your help
	Not what I expected, doesn't look like it would last long and I never got my supposedly bonus items. I even registered my spiralizer on their webwebsite and nothing no feed back about my bonus .you get what you pay for.
	Broke after a few uses
3	I only gave 3 stars because i returned it due to only getting one in the package of 2. Was not able to keep the one and have another shipped. so i returned it all together. Still not convinced that i want this material in a hot oven in contact with my food though.
	kind of leaky at the top and drips down the side some
	I should have listened to the other reviews that I read. I only had the cookware for about 3 months and I did everything that I was supposed to do when taking care of them and cooking in them. I used the correct sponge to wash them, bamboo utensils and still the black coating came off!! I wouldn't recommend these to anyone!! Cuisinart is supposed to be one of the top brands but they missed the mark with this one!!!!
4	This is one of the first home appliances I've purchased from Amazon, and while its hard to give a glowing review for something so mundane as a dehumidifier, I'm pleased with the way this one works. Our basement is basically a finished root cellar, in a 19th Century home. We have work supplies, clothing, some exercise equipment as well as some musical instruments stored down there, so keeping the humidity level down is a critical need to preserve what we store down there. The unit has a small footprint, takes up about the same space as a paper shredder. It's quiet, and in our cellar, which is probably no more than 700 sq ft, on the mid-level setting the unit fills up within about 48-72 hours, depending on recent weather (if it rains a lot, it fills up in less time). The catch basin has a drainage hose in the event your basement/cellar has a central drain, or a swivel handle on the inside for easy removal and dumping. The fill level also leave enough room from the top so that water does not accidentally spill when transporting it for dumping. Overall a smart design feature. I can't speak for the energy efficiency of the unit (though it seems to be rated well), but overall it does the job, and I am pleased with the product. I notice that the drier air in our cellar seems less musty then it did before I installed the unit.
	Like the mugs but the do crack easily and we have broken a couple. I hand wash I don't run these through the dish washer.
	I have been using these for about a month. They are very good, solid spatulas. Like others have said, it would be nice if the handles were a little longer, but I use them to cook in nonstick cookware, everyday.
5	Great. Love it. Especially the flat pan for French toast and pancakes. Cooks great and clean up is a cinch. Lightweight. Wonderful buy. Good price too.
	I use this almost every day. Works great for one cup of tea with no mess. I love how the infuser rests in it's little stainless steal holder. Absolutely no little puddles of tea on the counter when the infuser comes out of the tea.
	' I normally don't review items but this one is an exception to the rule. Please note, it was not purchased through Amazon but the merchant I went through doesn't have a location to review and I use Amazon quite frequently and I believe it may help others in search of a good packaging system. I have only used the unit now for fifteen packages of pecans in rather rapid succession. Works flawlessly. I have scrapped three other Food Saver brand machines prior to this VacMaster due to one thing or another. As a novel idea, this machine comes with an additional vacuum seal as well as an extra sealing element. Spare parts, that is great as neither of the replacements had that option nor product support when they cratered. Again this machine has already passed its successors and my hope is that it will remain a viable part

	of our vacuum packing processes for some time to come. I highly recommend this unit to anyone whom needs a serious packaging system. Pivot

3. The statistics of the rating (How many reviews received * ratings)
   - The number of reviews received 5 rating : 3124901
   - The number of reviews received 4 rating : 731748
   - The number of reviews received 3 rating : 349559
   - The number of reviews received 2 rating : 241953
   - The number of reviews received 1 rating : 426917

4. The number of reviews for each of these three classes (mapped to 1 / mapped 0 / discard reviews(=rating3))
   - The number of positive reviews(star_rating is 4 or 5) is 3856649
   - The number of negative reviews(star_rating is 1 or 2) is 668870
   - The number of neutral reviews(star_rating is 3) is 349559

5. The average length of the reviews in terms of character length in your dataset before and after cleaning
   - The average length of the reviews before cleaning : 212.788635
   - The average length of the reviews after cleaning : 204.786705

6. The average length of the reviews in terms of character length in your dataset before and after cleaning + preprocessing.
   - The average length of the reviews before cleaning and preprocessing : 212.788635
   - The average length of the reviews after cleaning and preprocessing : 125.995285

7. Print three sample reviews before and after data cleaning + preprocessing

#	Before data cleaning + preprocessing	After data cleaning + preprocessing
Sample 1	Very nice knives.	nice knife
Sample 2	I will only buy Kyocera knives from now on. Nothing cuts this thin and stays sharp as long. Great knife.	buy kyocera knife nothing cut thin stay sharp long great knife
Sample 3	The description says "Suitable for all type of surfaces" but it is not true! We ordered this item for our large family (we like to cook a lot) but this does not work as advertised! This pressure cooker does NOT work for induction stoves. I called the manufacturer (Magefesa) and they confirmed that it will not work. We are exchanging this pot for the Stainless Steel version that is	description say suitable type surface true ordered item large family like cook lot work advertised pressure cooker work induction stove called manufacturer magefesa confirmed work exchanging pot stainless steel version supposed work induction stove magefesa said quart pot work induction stove description amazon clearly state work induction stove thankfully

	supposed to work on induction stoves. Magefesa said that the 14.3 quart pot will work on induction stoves and the description here on Amazon clearly states that it will work on induction stoves. Thankfully Amazon has a great return process so they will send a truck to pick this up and they will send the other version in a couple days.	amazon great return process send truck pick send version couple day

8. Perceptron - Report Accuracy, Precision, Recall, and f1-score on both the training and testing split of datasets.

Perceptron	Accuracy	Precision	Recall	F1-Score
Train data	0.892631	0.878462	0.911621	0.894735
Test data	0.855275	0.839193	0.877417	0.857879

9. SVM - Report Accuracy, Precision, Recall, and f1-score on both the training and testing split of datasets.

SVM	Accuracy	Precision	Recall	F1-Score
Train data	0.931019	0.933611	0.928191	0.930893
Test data	0.898075	0.897889	0.897303	0.897596

10. Logistic Regression - Report Accuracy, Precision, Recall, and f1-score on both the training and testing split of datasets.

Logistic Regression	Accuracy	Precision	Recall	F1-Score
Train data	0.915362	0.920910	0.908974	0.914903
Test data	0.902475	0.905449	0.897856	0.901636

11. Multinomial Naïve Bayes - Report Accuracy, Precision, Recall, and f1-score on both the training and testing split of datasets.

Multinomial Naïve Bayes	Accuracy	Precision	Recall	F1-Score
Train data	0.894025	0.895741	0.892117	0.893926
Test data	0.876175	0.876712	0.874203	0.875456

Hee Ji Park

USC ID : 4090715830 / HW1 - CSCI544

In [1]:
```
# import package and libraries
import pandas as pd
import numpy as np
import nltk
nltk.download('wordnet')
import re
from bs4 import BeautifulSoup
```

```
[nltk_data] Downloading package wordnet to
[nltk_data]     C:₩Users₩gmlwl₩AppData₩Roaming₩nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
```

In [2]:
```
# pip install bs4 # in case you don't have it installed

# Dataset: https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Kitchen_v1_00.tsv.gz
```

Read Data

In [3]:
```
# unzip .gz file
import gzip
import shutil
with gzip.open('amazon_reviews_us_Kitchen_v1_00.tsv.gz', 'rb') as f_in:
    with open('amazon_reviews_us_Kitchen_v1_00.tsv', 'wb') as f_out:
        shutil.copyfileobj(f_in, f_out)
```

In [4]:
```
# read and make it to .csv file
import pandas as pd
dfs = pd.read_csv('amazon_reviews_us_Kitchen_v1_00.tsv', sep='₩t', chunksize=50, error_bad_lines=False)
for df in dfs:
    df.to_csv('amazon_reviews_us_Kitchen_v1_00.csv', sep=',', mode='a')
```

```
b'Skipping line 16148: expected 15 fields, saw 22₩n'
b'Skipping line 20100: expected 15 fields, saw 22₩n'
b'Skipping line 45178: expected 15 fields, saw 22₩n'
b'Skipping line 48700: expected 15 fields, saw 22₩n'
b'Skipping line 63331: expected 15 fields, saw 22₩n'
```

```
b'Skipping line 86053: expected 15 fields, saw 22₩n'
b'Skipping line 720217: expected 22 fields, saw 29₩n'
```

In [5]:
```
# read .csv file using pandas
df = pd.read_csv('amazon_reviews_us_Kitchen_v1_00.csv')
df.head()
```

Out[5]:

	Unnamed: 0	marketplace	customer_id	review_id	product_id	product_parent	product_title	product_category	star_rating	helpful_v
0	0.0	US	37000337	R3DT59XH7HXR9K	B00303FI0G	529320574	Arthur Court Paper Towel Holder	Kitchen	5	
1	1.0	US	15272914	R1LFS11BNASSU8	B00JCZKZN6	274237558	Olde Thompson Bavaria Glass Salt and Pepper Mi...	Kitchen	5	
2	2.0	US	36137863	R296RT05AG0AF6	B00JLIKA5C	544675303	Progressive International PL8 Professional Man...	Kitchen	5	
3	3.0	US	43311049	R3V37XDZ7ZCI3L	B000GBNB8G	491599489	Zyliss Jumbo Garlic Press	Kitchen	5	
4	4.0	US	13763148	R14GU232NQFYX2	B00VJ5KX9S	353790155	1 X Premier Pizza Cutter - Stainless Steel 14"...	Kitchen	5	

Keep Reviews and Ratings

In [6]:
```
# only keep the Reviews and Ratings fields in the input data frame to generate data.

rdf = df[['star_rating','review_body']]
rdf.head()
```

Out[6]:

	star_rating	review_body
0	5	Beautiful. Looks great on counter.
1	5	I personally have 5 days sets and have also bo...
2	5	Fabulous and worth every penny. Used for clean...
3	5	A must if you love garlic on tomato marinara s...
4	5	Worth every penny! Buy one now and be a pizza ...

Report #1 ----------

- Print three sample reviews along with corresponding ratings

In [7]:
```
# [sample review] star_rating - 1
rating1 = rdf[rdf.star_rating == '1'].sample(3).index

print("* Three sample review (rating = 1) ---------------------------------------- ")
for i in range(len(rating1)):
        print('{0}. {1}'.format(i+1, rdf.loc[rating1[i],'review_body']))
```

```
* Three sample review (rating = 1) ----------------------------------------
1. I bought this blender with high hopes, seeing all those informercials claiming that it blends so smoothly. Well when I pu
rchased it I was hoping to get a well blended fruit and vegetable drink, instead I got a chewy mess, and I feel taken for al
l the money that I paid for it. The 72 oz pitcher is a big waste of money. I can only use the single serve for a better blen
d. I wish I knew this before I purchased it, I could have paid way less for just a single serve, I will not be fooled again,
and I do not recommend to anybody!!! Nothing against amazon, just this blender.
2. What should I try to grill by using a palm sized set? Probably the worst product that I have bought from Amazon.
3. Both glasses were broken in the box - - very thin glass and not packed well. I was dissappointed.<br />I like the idea an
d also ordered teh frosted style, they did arrive unbroken but as I said, VERY THIN GLASS -- I don't know that they keep bee
r cold longer.
```

In [8]:
```
# [sample review] star_rating - 2
rating2 = rdf[rdf.star_rating == '2'].sample(3).index
print("* Three sample review (rating = 2) ---------------------------------------- ")
for i in range(len(rating2)):
        print('{0}. {1}'.format(i+1, rdf.loc[rating2[i],'review_body']))
```

```
* Three sample review (rating = 2) ----------------------------------------
1. Hi, I received my new coffeemaker, but the carafe lid was broked... How can I hve a new one on well conditions? I need to
order a new one?<br />Thanks for your help
2. Not what I expected, doesn't look like it would last long and I never got my supposedly bonus items. I even registered my
```

spiralizer on their webwebsite and nothing no feed back about my bonus .you get what you pay for.
3. Broke after a few uses

In [9]:
```
# [sample review] star_rating - 3
rating3 = rdf[rdf.star_rating == '3'].sample(3).index
print("* Three sample review (rating = 3) ---------------------------------------- ")
for i in range(len(rating3)):
        print('{0}. {1}'.format(i+1, rdf.loc[rating3[i],'review_body']))
```

```
* Three sample review (rating = 3) ----------------------------------------
1. I only gave 3 stars because i returned it due to only getting one in the package of 2. Was not able to keep the one and h
ave another shipped. so i returned it all together. Still not convinced that i want this material in a hot oven in contact w
ith my food though.
2. kind of leaky at the top and drips down the side some
3. I should have listened to the other reviews that I read. I only had the cookware for about 3 months and I did everything
that I was supposed to do when taking care of them and cooking in them. I used the correct sponge to wash them, bamboo utens
ils and still the black coating came off!! I wouldn't recommend these to anyone!! Cuisinart is supposed to be one of the top
brands but they missed the mark with this one!!!!
```

In [10]:
```
# [sample review] star_rating - 4
rating4 = rdf[rdf.star_rating == '4'].sample(3).index
print("* Three sample review (rating = 4) ---------------------------------------- ")
for i in range(len(rating4)):
        print('{0}. {1}'.format(i+1, rdf.loc[rating4[i],'review_body']))
```

```
* Three sample review (rating = 4) ----------------------------------------
1. This is one of the first home appliances I've purchased from Amazon, and while its hard to give a glowing review for some
thing so mundane as a dehumidifier, I'm pleased with the way this one works.  Our basement is basically a finished root ce
llar, in a 19th Century home.  We have work supplies, clothing, some exercise equipment as well as some musical instruments
stored down there, so keeping the humidity level down is a critical need to preserve what we store down there.  The unit h
as a small footprint, takes up about the same space as a paper shredder.  It's quiet, and in our cellar, which is probably n
o more than 700 sq ft, on the mid-level setting the unit fills up within about 48-72 hours, depending on recent weather (if
it rains a lot, it fills up in less time).  The catch basin has a drainage hose in the event your basement/cellar has a ce
ntral drain, or a swivel handle on the inside for easy removal and dumping.  The fill level also leave enough room from the
top so that water does not accidentally spill when transporting it for dumping.  Overall a smart design feature.  I can't
speak for the energy efficiency of the unit (though it seems to be rated well), but overall it does the job, and I am please
d with the product.  I notice that the drier air in our cellar seems less musty then it did before I installed the unit.
2. Like the mugs but the do crack easily and we have broken a couple.  I hand wash I don't run these through the dish washe
r.
3. I have been using these for about a month.  They are very good, solid spatulas.  Like others have said, it would be nice
if the handles were a little longer, but I use them to cook in nonstick cookware, everyday.
```

In [11]:
```
# [sample review] star_rating - 5
rating5 = rdf[rdf.star_rating == '5'].sample(3).index
print("* Three sample review (rating = 5) ---------------------------------------- ")
for i in range(len(rating5)):
        print('{0}. {1}'.format(i+1, rdf.loc[rating5[i],'review_body']))
```

```
* Three sample review (rating = 5) ----------------------------------------
1. Great.  Love it.  Especially the flat pan for French toast and pancakes.  Cooks great and clean up is a cinch.  Lightweig
ht.<br />Wonderful buy.  Good price too.
2. I use this almost every day.  Works great for one cup of tea with no mess.  I love how the infuser rests in it's little s
tainless steal holder.  Absolutely no little puddles of tea on the counter when the infuser comes out of the tea.
3. I normally don't review items but this one is an exception to the rule. Please note, it was not purchased through Amazon
but the merchant I went through doesn't have a location to review and I use Amazon quite frequently and I believe it may hel
p others in search of a good packaging system. I have only used the unit now for fifteen packages of pecans in rather rapid
succession. Works flawlessly. I have scrapped three other Food Saver brand machines prior to this VacMaster due to one thing
or another. As a novel idea, this machine comes with an additional vacuum seal as well as an extra sealing element. Spare pa
rts, that is great as neither of the replacements had that option nor product support when they cratered.<br /><br />Again t
his machine has already passed its successors and my hope is that it will remain a viable part of our vacuum packing process
es for some time to come. I highly recommend this unit to anyone whom needs a serious packaging system.<br /><br />Pivot
```

Labelling Reviews:

The reviews with rating 4,5 are labelled to be 1 and 1,2 are labelled as 0. Discard the reviews with rating 3'

In [12]:
```
import warnings
warnings.filterwarnings('ignore')

# check the value of the star_rating
rdf['star_rating'].unique()
```

Out[12]:
```
array(['5', '1', '3', '4', '2', 'star_rating', '5.0', '2.0', '3.0', '4.0',
       nan, '1.0'], dtype=object)
```

we can find the strange value like 'star_rating' and 'nan' value. In this case, we should remove these values at first.

In [13]:
```
# remove strange star_rating values like 'star_rating' & nan
rdf.drop(rdf[rdf['star_rating'].isna()].index, inplace = True)
rdf.drop(rdf[rdf['star_rating'] == 'star_rating'].index, inplace = True)

# change float type to int type in order to combine int value and float value
rdf.loc[:,'label'] =  rdf['star_rating'].astype(float).astype(int)
```

In [14]:
```
# Report the number of each grade(star_rating)

rating_count = {k: v for k, v in zip(rdf['label'].value_counts().index, rdf['label'].value_counts())}
print(rating_count)
```

```
{5: 3124901, 4: 731748, 1: 426917, 3: 349559, 2: 241953}
```

In [15]:
```
print("Report the statistics of the ratings ----------")
print("The number of reviews received 5 rating : ", rating_count[5])
print("The number of reviews received 4 rating : ", rating_count[4])
print("The number of reviews received 3 rating : ", rating_count[3])
print("The number of reviews received 2 rating : ", rating_count[2])
print("The number of reviews received 1 rating : ", rating_count[1])
```

```
Report the statistics of the ratings ----------
The number of reviews received 5 rating :  3124901
The number of reviews received 4 rating :  731748
The number of reviews received 3 rating :  349559
The number of reviews received 2 rating :  241953
The number of reviews received 1 rating :  426917
```

Report #2 ----------

- The number of reviews received 5 rating (star_rating is 5) is 3124901
- The number of reviews received 4 rating (star_rating is 4) is 731748
- The number of reviews received 3 rating (star_rating is 3) is 349559
- The number of reviews received 2 rating (star_rating is 2) is 241953
- The number of reviews received 1 rating (star_rating is 1) is 426917

In [16]:
```
# To create binary labels, mapping the ratings.

rdf.loc[(rdf['label'] < 3), 'label'] = 0
rdf.loc[(rdf['label'] == 3), 'label'] = -999
rdf.loc[(rdf['label'] > 3), 'label'] = 1
```

In [17]:
```
# Report the number of reviews for each of these three classes

rating_labelcount = {k: v for k, v in zip(rdf['label'].unique(), rdf['label'].value_counts())}
rating_labelcount
```

Out[17]:
```
{1: 3856649, 0: 668870, -999: 349559}
```

In [18]:
```
print("The number of positive reviews (mapped to 1 / star_rating is 4 or 5) : ", rating_labelcount[1])
print("The number of negative reviews (mapped to 0 / star_rating is 1 or 2) : ", rating_labelcount[0])
print("The number of neutral reviews (star_rating is 3) : ", rating_labelcount[-999])
```

```
The number of positive reviews (mapped to 1 / star_rating is 4 or 5) :  3856649
The number of negative reviews (mapped to 0 / star_rating is 1 or 2) :  668870
```

```
The number of neutral reviews (star_rating is 3) :  349559
```

Report #3 ----------

- The number of positive reviews(star_rating is 4 or 5) is 3856649
- The number of negative reviews(star_rating is 1 or 2) is 668870
- The number of neutral reviews(star_rating is 3) is 349559

In [19]:
```
# Keep only two classes: positive and negative
refined_rdf = rdf.loc[(rdf['label'] >= 0), ['review_body','label']]
refined_rdf.drop(refined_rdf[refined_rdf['review_body'].isna()].index, inplace = True)
refined_rdf.tail()
```

Out[19]:

	review_body	label
4972577	After a month of heavy use, primarily as a chi...	1
4972578	I've used my Le Creuset enameled cast iron coo...	1
4972579	According to my wife, this is \\"the best birt...	1
4972580	Hoffritz has a name of producing a trendy and ...	1
4972581	OK. I was late to snap to the Dead Reckoners. ...	1

We select 200000 reviews randomly with 100,000 positive and 100,000 negative reviews.

In [20]:
```
# select 200,000 reviews to perform the required tasks on the downsized datasets.

positive = refined_rdf[refined_rdf.label == 1][:100000]
negative = refined_rdf[refined_rdf.label == 0][:100000]
review = pd.concat([positive,negative],ignore_index=True)
```

2. Data Cleaning

In [21]:
```
# store three sample reviews in order to compare with reviews after data cleaning + preprocessing
sample1_before = review.review_body[10000]
```

```
sample2_before = review.review_body[20000]
sample3_before = review.review_body[100000]
```

Convert the all reviews into the lower case.

In [22]:
```
# convert the all reviews into the lower case
review["preprocess_review"] = review["review_body"].str.lower()
review.head()
```

Out[22]:

	review_body	label	preprocess_review
0	Beautiful. Looks great on counter.	1	beautiful. looks great on counter.
1	I personally have 5 days sets and have also bo...	1	i personally have 5 days sets and have also bo...
2	Fabulous and worth every penny. Used for clean...	1	fabulous and worth every penny. used for clean...
3	A must if you love garlic on tomato marinara s...	1	a must if you love garlic on tomato marinara s...
4	Worth every penny! Buy one now and be a pizza ...	1	worth every penny! buy one now and be a pizza ...

remove the HTML and URLs from the reviews

In [23]:
```
# remove HTML from the reviews
review["preprocess_review"] = review["preprocess_review"].apply(lambda x: BeautifulSoup(x).get_text())
```

In [24]:
```
# remove URLs from the reviews
review["preprocess_review"] = review["preprocess_review"].str.replace('http[s]?://(?:[a-zA-Z]|[0-9]|[$-_@.&+]|[!*W(W),]|(?:
```

remove non-alphabetical characters

In [25]:
```
# remove non-alphabetical characters
review.preprocess_review = review.preprocess_review.str.replace('[^a-zA-ZW'+W'+]', ' ')
review.preprocess_review
```

Out[25]:
```
0                      beautiful   looks great on counter
1         i personally have   days sets and have also bo...
2         fabulous and worth every penny  used for clean...
3         a must if you love garlic on tomato marinara s...
4         worth every penny  buy one now and be a pizza ...
5         awesome  first fryer i have owned and what a l...
6                           very good item  quick delivery
```

7	sharp and look great
8	my friend loves it
9	works as expected hold it close for quicker...
10	great
11	fantastic product love these pots
12	work great well worth the money and so much n...
13	holds a loooot of liquid bigger than it looks
14	the decobros k cup storage unit is very well m...
15	met my expectations
16	fits under my full sized keurig perfectly dr...
17	love it
18	love it
19	i love these after a lot of moves i've lost ...
20	this scoop has nice weight and feel in your h...
21	absolutely love my nutri ninja
22	i was having a hard time opening small lids on...
23	i've been waiting my whole life for these
24	i use this cast iron grill griddle for many th...
25	works great nice and slim fits right next to...
26	i have been wanting this for a while now and i...
27	works well easy to clean and fits well in the...
28	grinds coffee like a boss there is a static i...
29	spins through a sweet potato like it's nothing...
...	...
199970	excellent convection oven highly recommended...
199971	the brewed coffee will not stay hot in the pot...
199972	would not cut through a potato
199973	it doesn't work as great as i had hoped i ha...
199974	i bought this a couple months ago it leaks al...
199975	this appliance looks nice but cooks very uneve...
199976	well it worked about times very well then...
199977	horrible looks sort of nice out of the box wil...
199978	what is the world coming to when peeling a har...
199979	hots soups tend to build up extreme pressure c...
199980	it is smaller than it appeared pretty but wis...
199981	shoddy construction made out of plastic not t...
199982	very low quality
199983	can be fun for the first timer best way thou...
199984	the cupcakes came out crooked no matter what i...
199985	i returned it too narrow could only be use...
199986	waste of money and a pain to clean
199987	bad plastic taste affects the coffee in a big...
199988	broken when i received it returned
199989	this bottle totally disappointed me i didn't ...
199990	makes mush of vegetables not chopped
199991	we feel we wasted money on this it stays on f...
199992	this can opener is difficult to use i have to...
199993	first time i used this grinder the motor burnt...
199994	none of the three lids fit the shaker they we...

```
199995    i wish i could return this item for several re...
199996    i recently ordered   brentwood hb   stick blen...
199997    i bought two of these  they both leave my kniv...
199998    this device is very cumbersome and somewhat un...
199999    the pictures is not match the product    the p...
Name: preprocess_review, Length: 200000, dtype: object
```

Remove the extra spaces between the words

In [26]:
```
# remove the extra spaces between the words
review.preprocess_review = review.preprocess_review.replace('Ws+', ' ', regex=True)
review.preprocess_review
```

Out[26]:
```
0                         beautiful looks great on counter
1         i personally have days sets and have also boug...
2         fabulous and worth every penny used for cleani...
3         a must if you love garlic on tomato marinara s...
4         worth every penny buy one now and be a pizza s...
5         awesome first fryer i have owned and what a lu...
6                           very good item quick delivery
7                                    sharp and look great
8                                      my friend loves it
9         works as expected hold it close for quicker re...
10                                                  great
11                       fantastic product love these pots
12        work great well worth the money and so much ni...
13           holds a loooot of liquid bigger than it looks
14        the decobros k cup storage unit is very well m...
15                                    met my expectations
16        fits under my full sized keurig perfectly draw...
17                                                love it
18                                                love it
19        i love these after a lot of moves i've lost so...
20        this scoop has nice weight and feel in your ha...
21                          absolutely love my nutri ninja
22        i was having a hard time opening small lids on...
23               i've been waiting my whole life for these
24        i use this cast iron grill griddle for many th...
25        works great nice and slim fits right next to t...
26        i have been wanting this for a while now and i...
27        works well easy to clean and fits well in the ...
28        grinds coffee like a boss there is a static is...
29        spins through a sweet potato like it's nothing...
                                ...
199970    excellent convection oven highly recommended u...
199971    the brewed coffee will not stay hot in the pot...
199972                       would not cut through a potato
```

```
199973    it doesn't work as great as i had hoped i have...
199974    i bought this a couple months ago it leaks all...
199975    this appliance looks nice but cooks very uneve...
199976    well it worked about times very well then i pu...
199977    horrible looks sort of nice out of the box wil...
199978    what is the world coming to when peeling a har...
199979    hots soups tend to build up extreme pressure c...
199980    it is smaller than it appeared pretty but wish...
199981    shoddy construction made out of plastic not th...
199982                                    very low quality
199983    can be fun for the first timer best way though...
199984    the cupcakes came out crooked no matter what i...
199985    i returned it too narrow could only be used wi...
199986                  waste of money and a pain to clean
199987    bad plastic taste affects the coffee in a big ...
199988                  broken when i received it returned
199989    this bottle totally disappointed me i didn't l...
199990                makes mush of vegetables not chopped
199991    we feel we wasted money on this it stays on fa...
199992    this can opener is difficult to use i have to ...
199993    first time i used this grinder the motor burnt...
199994    none of the three lids fit the shaker they wer...
199995    i wish i could return this item for several re...
199996    i recently ordered brentwood hb stick blenders...
199997    i bought two of these they both leave my knive...
199998    this device is very cumbersome and somewhat un...
199999    the pictures is not match the product the pict...
Name: preprocess_review, Length: 200000, dtype: object
```

perform contractions on the reviews.

In [27]:
```
# perform contractions on the reviews
import contractions

def contractionfunction(s):
    s = s.apply(lambda x: contractions.fix(x))
    return s
```

In [28]:
```
review.preprocess_review = contractionfunction(review.preprocess_review)
```

In [29]:
```
review.preprocess_review
```

Out[29]:
```
0                         beautiful looks great on counter
1         i personally have days sets and have also boug...
2         fabulous and worth every penny used for cleani...
3         a must if you love garlic on tomato marinara s...
```

4	worth every penny buy one now and be a pizza s...
5	awesome first fryer i have owned and what a lu...
6	very good item quick delivery
7	sharp and look great
8	my friend loves it
9	works as expected hold it close for quicker re...
10	great
11	fantastic product love these pots
12	work great well worth the money and so much ni...
13	holds a loooot of liquid bigger than it looks
14	the decobros k cup storage unit is very well m...
15	met my expectations
16	fits under my full sized keurig perfectly draw...
17	love it
18	love it
19	i love these after a lot of moves I have lost ...
20	this scoop has nice weight and feel in your ha...
21	absolutely love my nutri ninja
22	i was having a hard time opening small lids on...
23	I have been waiting my whole life for these
24	i use this cast iron grill griddle for many th...
25	works great nice and slim fits right next to t...
26	i have been wanting this for a while now and i...
27	works well easy to clean and fits well in the ...
28	grinds coffee like a boss there is a static is...
29	spins through a sweet potato like it is nothin...
	...
199970	excellent convection oven highly recommended u...
199971	the brewed coffee will not stay hot in the pot...
199972	would not cut through a potato
199973	it does not work as great as i had hoped i hav...
199974	i bought this a couple months ago it leaks all...
199975	this appliance looks nice but cooks very uneve...
199976	well it worked about times very well then i pu...
199977	horrible looks sort of nice out of the box wil...
199978	what is the world coming to when peeling a har...
199979	hots soups tend to build up extreme pressure c...
199980	it is smaller than it appeared pretty but wish...
199981	shoddy construction made out of plastic not th...
199982	very low quality
199983	can be fun for the first timer best way though...
199984	the cupcakes came out crooked no matter what i...
199985	i returned it too narrow could only be used wi...
199986	waste of money and a pain to clean
199987	bad plastic taste affects the coffee in a big ...
199988	broken when i received it returned
199989	this bottle totally disappointed me i did not ...
199990	makes mush of vegetables not chopped
199991	we feel we wasted money on this it stays on fa...

```
199992    this can opener is difficult to use i have to ...
199993    first time i used this grinder the motor burnt...
199994    none of the three lids fit the shaker they wer...
199995    i wish i could return this item for several re...
199996    i recently ordered brentwood hb stick blenders...
199997    i bought two of these they both leave my knive...
199998    this device is very cumbersome and somewhat un...
199999    the pictures is not match the product the pict...
Name: preprocess_review, Length: 200000, dtype: object
```

In [30]:
```
## For report, calculate the length of the reviews before cleanging.
length_sum = 0
for text in review.review_body:
    length_sum += len(text)

average_length_reviews = length_sum / 200000

print("Compare to the average length of the reviews before and after cleaning ----------")
print("The average length of the reviews before cleaning : ", average_length_reviews)
```

```
Compare to the average length of the reviews before and after cleaning ----------
The average length of the reviews before cleaning :  212.788635
```

In [31]:
```
## For report, calculate the length of the reviews after cleaning
length_sum = 0
for text in review.preprocess_review:
    length_sum += len(text)

average_length_reviews = length_sum / 200000
print("The average length of the reviews after cleaning : ", average_length_reviews)
```

```
The average length of the reviews after cleaning :  204.786705
```

Report #4 ----------

- The average length of the reviews in terms of character length in your dataset before and after cleaning
- The average length of the reviews before cleaning : 212.788635
- The average length of the reviews after cleaning : 204.786705

---

3. Preprocessing

remove the stop words

In [32]:
```
# remove the stop words using NLTK package
from nltk.corpus import stopwords
nltk.download('stopwords')
stop_words = (set(stopwords.words("english")) )

def removeStop(s):
    s_list = s.split()
    final_list = [word for word in s_list if word not in stop_words]
    final_string = ' '.join(final_list)
    return final_string
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:₩Users₩gmlwl₩AppData₩Roaming₩nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

In [33]:
```
review.preprocess_review = review.preprocess_review.apply(lambda x: removeStop(x))
```

perform lemmatization

In [34]:
```
# perform lemmatization
from nltk.stem import WordNetLemmatizer

def lemmatization_function(s):
    s_list = s.split()
    wordnet_lemmatizer = WordNetLemmatizer()
    final_list = [wordnet_lemmatizer.lemmatize(x) for x in s_list]
    final_string = ' '.join(final_list)
    return final_string
```

In [35]:
```
review.preprocess_review = review.preprocess_review.apply(lambda x: lemmatization_function(x))
```

In [36]:
```
# store three sample reviews in order to compare with reviews before data cleaning + preprocessing
sample1_after = review.preprocess_review[10000]
sample2_after = review.preprocess_review[20000]
sample3_after = review.preprocess_review[100000]
```

In [37]:
```
## For report, calculate the length of the reviews before cleanging + preprocessing.
length_sum = 0
for text in review.review_body:
    length_sum += len(text)
```

```
average_length_reviews = length_sum / 200000

print("Compare to the average length of the reviews before and after cleaning & preprocessing ----------")
print("The average length of the reviews before cleaning and preprocessing : ", average_length_reviews)
```

```
Compare to the average length of the reviews before and after cleaning & preprocessing ----------
The average length of the reviews before cleaning and preprocessing :  212.788635
```

In [38]:
```
## For report, calculate the length of the reviews after cleanging + preprocessing.
length_sum = 0
for text in review.preprocess_review:
    length_sum += len(text)

average_length_reviews = length_sum / 200000
print("The average length of the reviews after cleaning and preprocessing : ", average_length_reviews)
```

```
The average length of the reviews after cleaning and preprocessing :  125.995285
```

Report #5 ----------

- The average length of the reviews before cleaning and preprocessing : 212.788635
- The average length of the reviews after cleaning and preprocessing : 125.995285

Report #6 ----------

In [39]:
```
print("Three sample reviews before and after data cleaning + preprocessing ----------")
print("- Sample 1 before data cleaning + preprocessing : ₩n", sample1_before)
print("- Sample 1 after data cleaning + preprocessing : ₩n", sample1_after)
print("- Sample 2 before data cleaning + preprocessing : ₩n", sample2_before)
print("- Sample 2 after data cleaning + preprocessing : ₩n", sample2_after)
print("- Sample 3 before data cleaning + preprocessing : ₩n", sample3_before)
print("- Sample 3 after data cleaning + preprocessing : ₩n", sample3_after)
```

```
Three sample reviews before and after data cleaning + preprocessing ----------
- Sample 1 before data cleaning + preprocessing :
 Very nice knives.
- Sample 1 after data cleaning + preprocessing :
 nice knife
- Sample 2 before data cleaning + preprocessing :
 I will only buy Kyocera knives from now on.  Nothing cuts this thin and stays sharp as long.  Great knife.
- Sample 2 after data cleaning + preprocessing :
```

```
 buy kyocera knife nothing cut thin stay sharp long great knife
- Sample 3 before data cleaning + preprocessing :
 The description says &#34;Suitable for all type of surfaces&#34; but it is not true!<br /><br />We ordered this item for ou
r large family (we like to cook a lot) but this does not work as advertised!<br /><br />This pressure cooker does NOT work f
or induction stoves.  I called the manufacturer (Magefesa) and they confirmed that it will not work. We are exchanging this
pot for the Stainless Steel version that is supposed to work on induction stoves.  Magefesa said that the 14.3 quart pot wil
l work on induction stoves and the description here on Amazon clearly states that it will work on induction stoves.<br /><br
/>Thankfully Amazon has a great return process so they will send a truck to pick this up and they will send the other versio
n in a couple days.
- Sample 3 after data cleaning + preprocessing :
 description say suitable type surface true ordered item large family like cook lot work advertised pressure cooker work ind
uction stove called manufacturer magefesa confirmed work exchanging pot stainless steel version supposed work induction stov
e magefesa said quart pot work induction stove description amazon clearly state work induction stove thankfully amazon great
return process send truck pick send version couple day
```

Split dataset into 80% training dataset and 20% testing dataset

In [40]:
```
from sklearn.model_selection import train_test_split

x_train, x_test, y_train, y_test = train_test_split(review['preprocess_review'], review['label'], test_size=0.2, random_sta
```

TF-IDF Feature Extraction

In [41]:
```
from sklearn.feature_extraction.text import TfidfVectorizer

tfvector = TfidfVectorizer()
tf_x_train = tfvector.fit_transform(x_train)
tf_x_test = tfvector.transform(x_test)
```

Report #7 ----------

Perceptron

In [42]:
```
from sklearn.metrics import precision_score, recall_score, f1_score, accuracy_score
from sklearn.linear_model import Perceptron

pct = Perceptron(tol=1e-3, random_state=0)
pct.fit(tf_x_train, y_train)
```

```
pct_y_train_pred = pct.predict(tf_x_train)
pct_y_test_pred = pct.predict(tf_x_test)
```

In [43]:
```
print('Perceptron_Train_Accuracy: %f' % accuracy_score(y_train, pct_y_train_pred))
print('Perceptron_Train_Precision: %f' % precision_score(y_train, pct_y_train_pred))
print('Perceptron_Train_Recall: %f' % recall_score(y_train, pct_y_train_pred))
print('Perceptron_Train_F1 Score: %f' % f1_score(y_train, pct_y_train_pred))

print('Perceptron_Test_Accuracy: %f' % accuracy_score(y_test, pct_y_test_pred))
print('Perceptron_Test_Precision: %f' % precision_score(y_test, pct_y_test_pred))
print('Perceptron_Test_Recall: %f' % recall_score(y_test, pct_y_test_pred))
print('Perceptron_Test_F1 Score: %f' % f1_score(y_test, pct_y_test_pred))
```

```
Perceptron_Train_Accuracy: 0.892631
Perceptron_Train_Precision: 0.878462
Perceptron_Train_Recall: 0.911621
Perceptron_Train_F1 Score: 0.894735
Perceptron_Test_Accuracy: 0.855275
Perceptron_Test_Precision: 0.839193
Perceptron_Test_Recall: 0.877417
Perceptron_Test_F1 Score: 0.857879
```

Report #8 ----------

SVM

In [44]:
```
from sklearn.svm import LinearSVC

svm = LinearSVC(random_state=0)
svm.fit(tf_x_train, y_train)
svm_y_train_pred = svm.predict(tf_x_train)
svm_y_test_pred = svm.predict(tf_x_test)
```

In [45]:
```
print('SVM_Train_Accuracy: %f' % accuracy_score(y_train, svm_y_train_pred))
print('SVM_Train_Precision: %f' % precision_score(y_train, svm_y_train_pred))
print('SVM_Train_Recall: %f' % recall_score(y_train, svm_y_train_pred))
print('SVM_Train_F1 Score: %f' % f1_score(y_train, svm_y_train_pred))

print('SVM_Test_Accuracy: %f' % accuracy_score(y_test, svm_y_test_pred))
print('SVM_Test_Precision: %f' % precision_score(y_test, svm_y_test_pred))
print('SVM_Test_Recall: %f' % recall_score(y_test, svm_y_test_pred))
print('SVM_Test_F1 Score: %f' % f1_score(y_test, svm_y_test_pred))
```

```
SVM_Train_Accuracy: 0.931019
SVM_Train_Precision: 0.933611
SVM_Train_Recall: 0.928191
SVM_Train_F1 Score: 0.930893
SVM_Test_Accuracy: 0.898075
SVM_Test_Precision: 0.897889
SVM_Test_Recall: 0.897303
SVM_Test_F1 Score: 0.897596
```

Report #9 ----------

Logistic Regression

In [46]:
```
from sklearn.linear_model import LogisticRegression

lr = LogisticRegression()
lr.fit(tf_x_train,y_train)
lr_y_train_pred = lr.predict(tf_x_train)
lr_y_test_pred = lr.predict(tf_x_test)
```

In [47]:
```
print('Logistic Regression_Train_Accuracy: %f' % accuracy_score(y_train, lr_y_train_pred))
print('Logistic Regression_Train_Precision: %f' % precision_score(y_train, lr_y_train_pred))
print('Logistic Regression_Train_Recall: %f' % recall_score(y_train, lr_y_train_pred))
print('Logistic Regression_Train_F1 Score: %f' % f1_score(y_train, lr_y_train_pred))

print('Logistic Regression_Test_Accuracy: %f' % accuracy_score(y_test, lr_y_test_pred))
print('Logistic Regression_Test_Precision: %f' % precision_score(y_test, lr_y_test_pred))
print('Logistic Regression_Test_Recall: %f' % recall_score(y_test, lr_y_test_pred))
print('Logistic Regression_Test_F1 Score: %f' % f1_score(y_test, lr_y_test_pred))
```

```
Logistic Regression_Train_Accuracy: 0.915362
Logistic Regression_Train_Precision: 0.920910
Logistic Regression_Train_Recall: 0.908974
Logistic Regression_Train_F1 Score: 0.914903
Logistic Regression_Test_Accuracy: 0.902475
Logistic Regression_Test_Precision: 0.905449
Logistic Regression_Test_Recall: 0.897856
Logistic Regression_Test_F1 Score: 0.901636
```

Report #10 ----------

Naive Bayes

In [48]:
```
from sklearn.naive_bayes import MultinomialNB
mnb = MultinomialNB()
mnb.fit(tf_x_train,y_train)
mnb_y_train_pred = mnb.predict(tf_x_train)
mnb_y_test_pred = mnb.predict(tf_x_test)
```

In [49]:
```
print('Naive Bayes_Train_Accuracy: %f' % accuracy_score(y_train, mnb_y_train_pred))
print('Naive Bayes_Train_Precision: %f' % precision_score(y_train, mnb_y_train_pred))
print('Naive Bayes_Train_Recall: %f' % recall_score(y_train, mnb_y_train_pred))
print('Naive Bayes_Train_F1 Score: %f' % f1_score(y_train, mnb_y_train_pred))
print('Naive Bayes_Test_Accuracy: %f' % accuracy_score(y_test, mnb_y_test_pred))
print('Naive Bayes_Test_Precision: %f' % precision_score(y_test, mnb_y_test_pred))
print('Naive Bayes_Test_Recall: %f' % recall_score(y_test, mnb_y_test_pred))
print('Naive Bayes_Test_F1 Score: %f' % f1_score(y_test, mnb_y_test_pred))
```

```
Naive Bayes_Train_Accuracy: 0.894025
Naive Bayes_Train_Precision: 0.895741
Naive Bayes_Train_Recall: 0.892117
Naive Bayes_Train_F1 Score: 0.893926
Naive Bayes_Test_Accuracy: 0.876175
Naive Bayes_Test_Precision: 0.876712
Naive Bayes_Test_Recall: 0.874203
Naive Bayes_Test_F1 Score: 0.875456
```