

AUTOATED REVIEW RATING SYSTEM

1. Project Overview

The Automated Review Rating System is a machine learning and NLP-based project that predicts star ratings (1–5) from customer reviews. It processes and analyzes review text to classify sentiment automatically, helping businesses save time, gain insights, and improve decision-making.

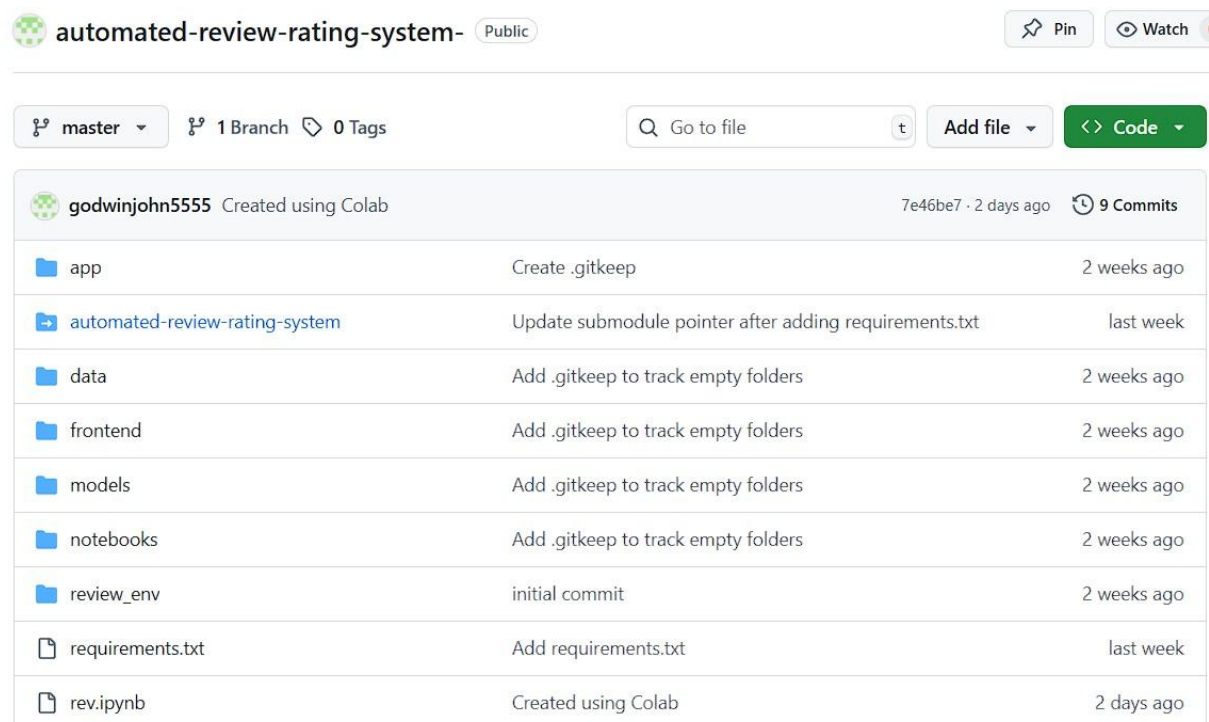
2. Environment Setup

- Python 3.12
- Libraries: pandas,nltk,scitkit-learn,numpy
- NLTK : Stopwords,punkt
- IDE: Google Colab,vs code

3. GitHub Project Setup

Created GitHub repository: automated-review-rating-system

Structure of directory



The screenshot shows the GitHub interface for a repository named 'automated-review-rating-system' by user 'godwinjohn5555'. The repository is public and was created using Colab. The commit history table is as follows:

Commit Message	Commit Hash	Time Ago
7e46be7 · 2 days ago	9 Commits	
app	Create .gitkeep	2 weeks ago
automated-review-rating-system	Update submodule pointer after adding requirements.txt	last week
data	Add .gitkeep to track empty folders	2 weeks ago
frontend	Add .gitkeep to track empty folders	2 weeks ago
models	Add .gitkeep to track empty folders	2 weeks ago
notebooks	Add .gitkeep to track empty folders	2 weeks ago
review_env	initial commit	2 weeks ago
requirements.txt	Add requirements.txt	last week
rev.ipynb	Created using Colab	2 days ago

4.Data.Collection

- Data was collected from Kaggle and it related Amazon Fine Food Reviews Dataset, a publicly available dataset containing over 568,000 customer reviews of food products from Amazon.
- Dataset Size; Total Records: 568,454 reviews, Columns: **10**
- Dataset link: <https://www.kaggle.com/snap/amazon-fine-foodreviews/downloads/amazon-fine-food-reviews.zip>
- Final dataset contains 2 column with Review and Rating
- 1star=52264, 2star=29743, 3star=42638, 4star=80654, 5star=363102

4.1.Balanced dataset

- Link for balanced dataset
=https://github.com/godwinjohn5555/automatedreview-rating-system-
- Balanced dataset was created with 5000 rows each rating

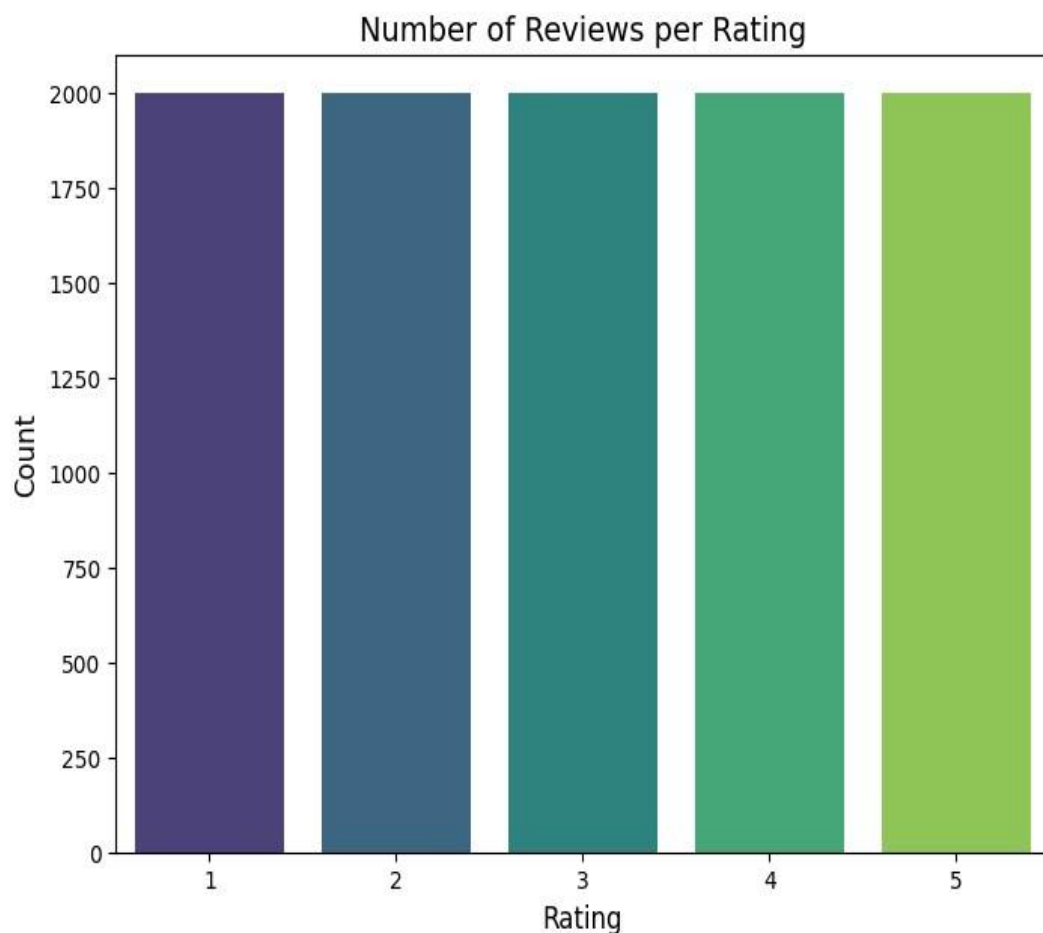


Fig count plot

Distribution of Reviews per Rating

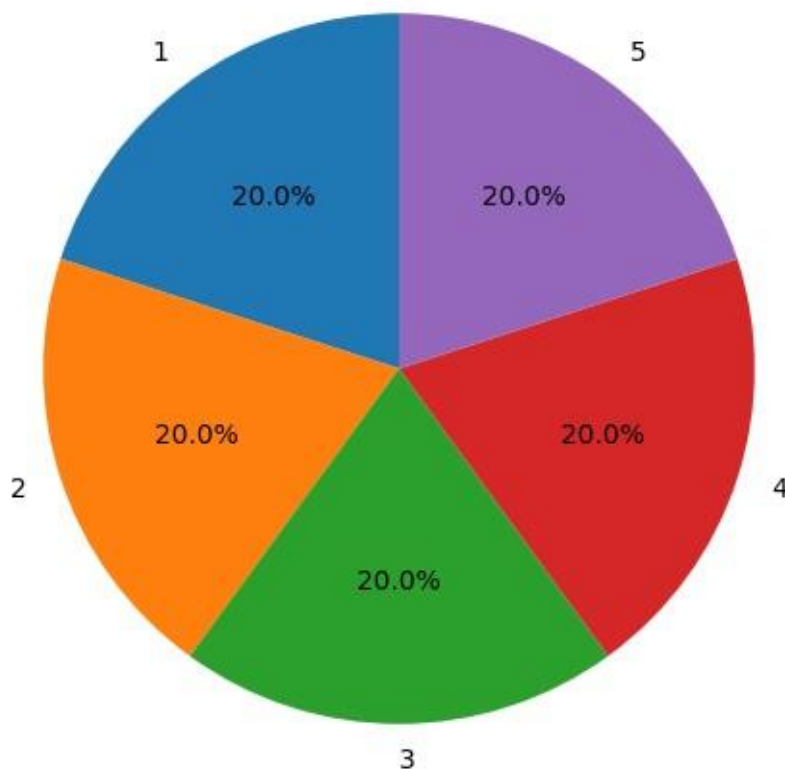


Fig pie chart

5.Data Preprocessing

Effective data preprocessing is essential to improve the performance and accuracy of machine learning models. The following techniques were applied to prepare the raw dataset for modeling.

5.1 Removing Duplicates

Duplicate rows containing the exact same review and rating were removed to prevent bias and overfitting. This ensured that the dataset only included unique observation.

Code:

```
review_cleaned = review.drop_duplicates(subset=['Review_text'], keep='first')  
print("shape before cleaning: ", review.shape)  
print("shape after cleaning: ", review_cleaned.shape)
```

5.2 Removing Conflicting Reviews

Some reviews had identical text but different start ratings. These inconsistencies can confuse the model. Such conflicting entries were identified and removed to maintain label clarity.

Code:

```
rating_nunique = rev.groupby('Text')['Rating'].nunique()
conflict_texts = set(rating_nunique[rating_nunique > 1].index)

before = rev.shape[0]
rev = rev[~rev['Text'].isin(conflict_texts)].copy()
print(f"Removed conflicts (same text, different ratings): {before - rev.shape[0]} rows; now {rev.shape[0]}")
```

5.3 Handling Missing Values

Rows with missing or null values particularly in the review text or rating columns were removed. This step ensured the dataset was complete and meaningful for analysts.

Code:

```
rev = rev.dropna(subset=["Review_text", "Rating"])
```

5.4 Dropping Unnecessary Columns

Non-essential columns such as review IDs, timestamps, or user identifiers were dropped. These fields did not contribute to the model and could introduce noise or privacy concerns.

5.5 Lowercasting Text

All interview text was converted to lowercase to maintain uniformity. This helps prevent duplication of tokens like "good" being treated as separate words.

5.6 Removing URLs

URLs present in the review text were removed using regular expressions.

5.7 Removing Emojis and Special Characters

It is the process of cleaning text by eliminating unnecessary symbols, emojis, and non-alphanumeric characters. This step ensures that the dataset contains only meaningful words, making it easier for NLP models to analyze and learn from the text.

5.8 Removing Punctuation

Removing punctuation means cleaning text by eliminating symbols such as .,!?;>()[] etc. These characters do not usually add meaning for sentiment analysis or review rating tasks, so removing them helps in simplifying the text and focusing on the actual words.

Code:

```
def remove_punctuation(text):  
    return text.translate(str.maketrans("", "", string.punctuation))  
review["Review_text"] = review["Review_text"].astype(str).apply(remove_punctuation)
```

5.9 Stopwords Removal

Stopwords are commonly used words in a language (such as *"is"*, *"the"*, *"in"*, *"on"*) that carry little meaning on their own. In NLP tasks, they are often removed to reduce noise and focus on the words that contribute most to the sentiment or meaning of the review.

We use the Natural Language Toolkit (NLTK) to access the English stopwords list. NLTK provides a predefined list of 179 stopwords.

Stopword Statistics

- Total Stopwords in NLTK (English): 179
- Total Stopwords Found in Dataset: 1,034,759

Top 10 Most Frequent Stopwords in Reviews

Stopword Count

the	94,151
i	70,251
a	55,741
and	55,104
to	48,442
of	40,009
it	38,371
is	32,946
this	29,954
in	24,913

Code:

```
Def remove_stopwords(text):
```

```
    words = text.split()    filtered = [word for word in words if  
word.lower() not in stop_words]    return "" ,join(filtered)
```

```
review["Review_text"] =
```

```
review["Review_text"].astype(str).apply(remove_stopwords)
```

5.10 Lemmatization

Lemmatization is the process of reducing words to their base or dictionary form (lemma) while keeping the meaning intact. For example:

- “running” → “run”
- “better” → “good”

It helps standardize words so NLP models can treat different forms of a word as the same.

Why lemmatization is better than stemming?

- Produces valid words – Lemmatization returns proper dictionary words, while stemming may give non-words (e.g., “studies → study” vs “studies → studi”).
- Cleaner text – Creates consistent, meaningful tokens for NLP models.

5.11 Filtering by Wordcounts

Filtering by word counts means removing reviews or text entries that are too short or too long, since they may not provide useful information for analysis.

6. Data Visualisation

Box plot

A box plot (or whisker plot) is a graphical representation of a dataset's distribution and variability. It shows the minimum, first quartile (Q1), median, third quartile (Q3), and maximum values, and highlights outliers. Box plots are useful for quickly understanding the spread, skewness, and presence of extreme values in the data.

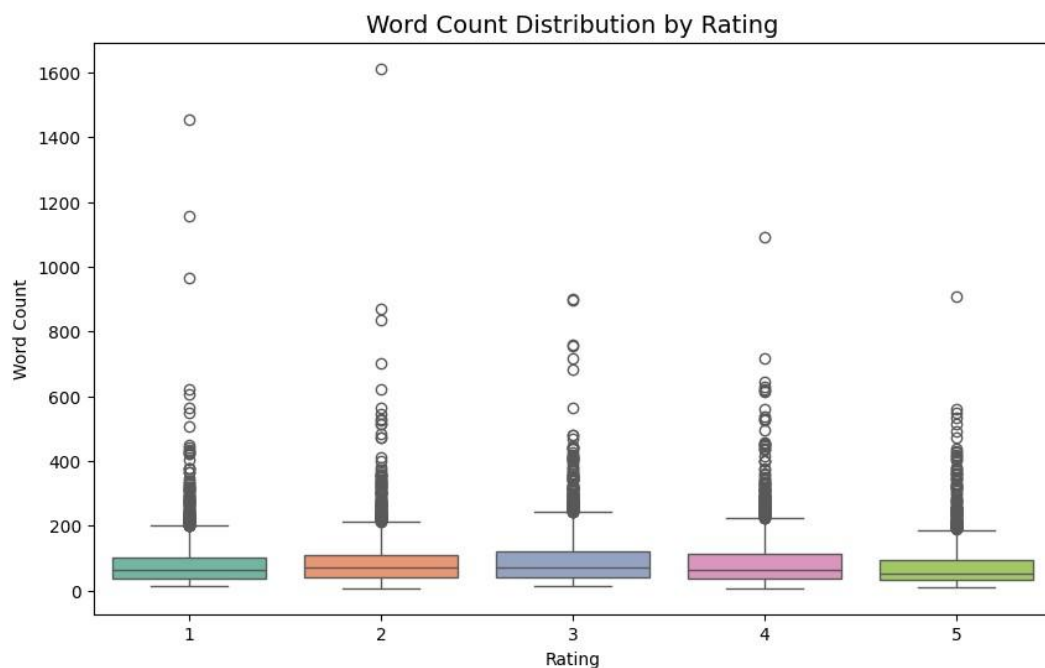


Fig box plot

Histogram

A histogram is a graphical representation of the distribution of numerical data.

It divides the data into intervals (bins) and shows the frequency of values within each bin. Histograms are useful for visualizing patterns, such as skewness, spread, and the shape of the dataset.

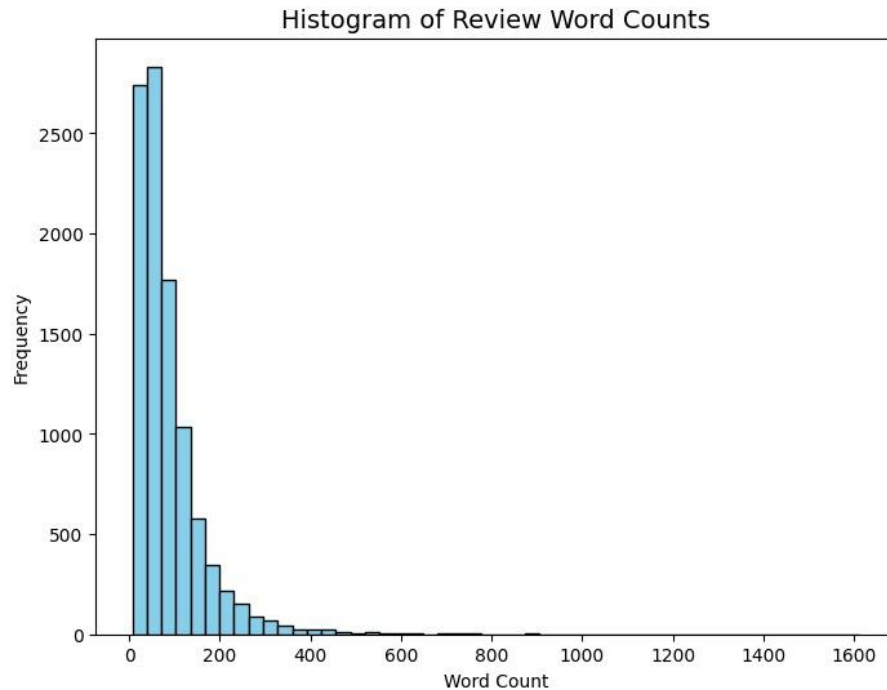


Fig histogram

Barplot

A bar plot is a type of chart used to represent categorical data with rectangular bars, where the length or height of each bar corresponds to the value or frequency of the category it represents. It provides a simple way to compare quantities across different groups or categories. The categories are usually placed on one axis, while the values are placed on the other, making it easy to visually compare differences between groups. Bar plots can be drawn vertically or horizontally depending on the type of comparison required.

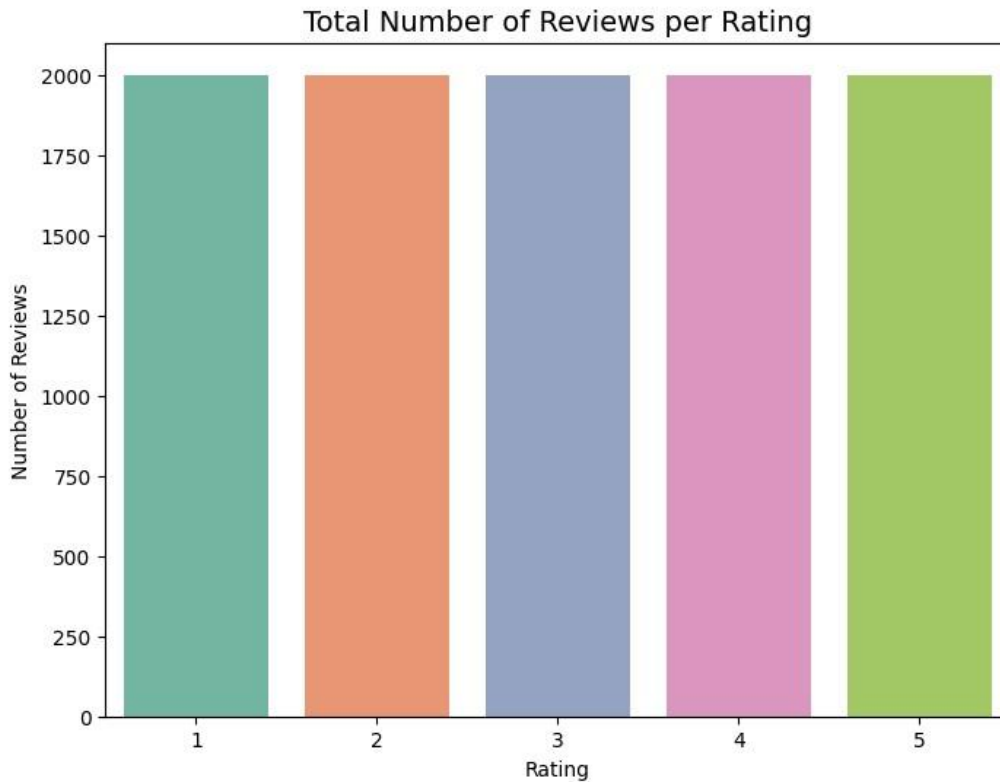


Fig barplot

Samples of 1 star rating

Showing 5 sample review for rating 1:

1. i bought this to make coconut milk for my coffee but it doesn't turn into milk even in a blender the bits are too big made me choke on my coffee will never reorder.
2. the tea i received is not caffeine free the box is different from the pictured box as well it has ok flavor but could be stronger per bag.
3. used of the bottles over the past month and have zero effect on helping with sleep it was no different than taking nothing expensive and worthless.
4. arrived with one package that was loose the other one tight had tape on it so i assume it was like that when it left outside packaging was fine would not order it again.
5. got it for a friend so naturally i didn't open it to look everything was thin and cheap save yourself money get pancake mix and a real pitcher the cake ring can be found at a thrift store.

Samples of 2 star rating

Showing 5 sample review for rating 2:

1. this is an attractive machine looks good in my kitchen my husband and i use different coffees and love the convenience of having whatever brew we need available cup by cup it s easy to operate the coffee smells good and tastes good having said that here s the downside it s not very well made it has been recalled for some electrical difficulty that could have been disastrous and the replacement broke down within months i am now on my fourth or fifth one and if this one goes bad in the next two years or so i am going shopping for another brand sometimes i think i keep buying them just because it s so fast efficient and pleasant to deal with amazon.
2. usually i buy freeze dried pears from whole foods but it costs an arm and a leg so i was super excited to see this available at a decent price disappointment doesn t even come close the pears weren t crunchy in the slightest bit more like stale and chewy in such a way that it brought about a slight ick factor i was expecting the same quality as whole foods my mistake but when i opened the bag it looked like someone had taken a pear slice and crumbled it like a paper ball and dried it we got this to have healthy snacks on hand but no one in my house can get over that slight ick taste and no one will eat it whatta waste.
3. this earl grey tea is just ok the bergamont flavor is muted and dull which caught me by surprise i was expecting better from this product as it is marketed and packaged as high end also the tea leaves are broken and not full leaves.
4. while these do have a great ingredient label the texture and flavor are imo off putting i bought the strawberry but it tastes like a cross between apple and pear and leaves a strange taste in my mouth but the texture is particularly odd for years i made my own fruit leather guess i ll have to back to it i really don t like writing negative reviews but if i had really paid attention to the poor reviews here i probably wouldn t have bought this i just hope i can save someone else from making the same mistake.
5. as a note to anyone buying this product i emailed watkins to ask about how the product is processed and received the following reply watkins original grapeseed oil fl oz is processed with hexane after researching hexane extraction there may still be slight traces of hexane in this product as in any

hexane processed product hexane is banned or restricted in the european union for cosmetics and the deep skin cosmetics database states research shows there may be toxicity problems just a note to those looking to buy that this may not be all natural common problem for most grapeseed oils.

Samples of 3 star rating

Showing 5 sample review for rating 3:

1. absolutely love the different flavors of coffee that were in this bundle very delicious and at a decent price i honestly was a little surprised that it came in such a big box only for the k cups to be in a brown bag and not in their own box like you buy at the store other than that small quirk i probably will buy my k cups from here again but not after looking at other sellers.
2. my family and i recently tasted this switch drink as well as the black cherry flavor i found the kiwi strawberry taste to not be very distinguishable ie without reading the can i would have had a hard time guessing what the flavor was my husband thought this was better than the cherry flavor as did the kids ages we all found this drink to be very sweet even without the added sugars i was not impressed with this drink and would not buy it again.
3. not thrilled with these doubt i will buy them again lid doesn't fit very well messy and did not make a good cup of coffee cost including shipping to me was filter cup plus coffee don't think it is worth it as i can buy a green mountain k cup for around cup and they are perfect every time.
4. this tea has a gentle laxative effect when i take it in the evening i spend much of the next day making trips to the toilet but nothing too urgent that said everyone's system is different i never take laxatives so i can't compare them to the tea unfortunately this chocolate flavor is faint and rather chalky i like regular smooth move better.
5. this is the best hot cocoa i have found so far but i hope i can find better it's not real chocolately wish it were a bit sweeter but on the whole i would buy it again if i can't find better.

Samples of 4 star rating

Showing 5 sample review for rating 4:

1. my dog loves these cookies i just wanted to let people know the box is oz and not oz as they are saying one ounce is no big deal so this is just an fyi no problem with amazon at all i use them frequently and this is the first time this has happened.

2. the curry arrived well packaged and timely and most importantly was delicious i highly recommend this company for future purchases.

3. the pieces of fruit were ripe flavorful firm not water logged and mushy like some canned fruit i ve tried the light syrup is only fruit juice not the typical light or heavy corn syrup which i was quite grateful to find the reason i did not give stars is there is an approximate to ratio of fruit to juice it is listed as servings but it was one cereal sized bowl for me including juice granted i should have made it servings considering the calories but to make servings you would need to use the small custard bowls which would not satiate me i wish they had an organic version yes i ve tried native forrest and i thought it was mushy i will reorder until i find another organic version i like i have bought other roland organic and non organic products that i have enjoyed so i consider this brand trustworthy.

4. why oh why do studios stamp deluxe edition on movies and then provide nothing to make it worthy of that title i absolutely love the movie beetlejuice and just about anything else tim burton has done the blu ray version of this movie looks and sounds better than any version i ve seen before the colors are sharp and the sound is greatly improved over other dvd iterations i have seen before the movie comes with a nice slipcover featuring the cover art in lenticular motion that s about the only really special feature you are going to get on this disc with it being the th anniv of this film i thought that surely something would have been included about the making of the film but there is nothing there is no commentary track no making of doc no interviews with the cast nothing oh you do get episodes of the beetlejuice cartoon yaawwwwnn and the original theatrical trailer i m not complaining too much because i do love this movie and it is worth the price just to see it in high definition it just ticks me off a little when studios stamp the deluxe edition on the cover when there is nothing deluxe about it.

5. one of the reasons i like this product is there are no added sweeteners just whole grains nuts and fruit i add additional sunflower seeds ground flax seed oat bran and milk to the muesli and heat it in the microwave for a minute or two i then add a spoonful of yogurt usually the fruit flavored type for a nice easy breakfast cereal.

Samples of 5 star rating

Showing 5 sample review for rating 5:

1. i bought these at a meijer s in blacklick reynoldsburg oh while visiting my in laws then while at the whole foods on the other side of town bought some fresh ground honey peanut butter i had to hide these pretzels because everyone was eating them saying they were the best they had ever had they truly are wonderful
2. crackers are great they come in a small enough package so they don t get stale either but i can t seem to put them down until they are almost gone try them
3. i bought this because my boyfriend likes green mountain coffee but had not tried this blend before he loved it so much he requested i buy it every time we get new k cups it is perfect for breakfast wake up and even smells good to me a non drinker of coffee
4. excellent coffee we blend regular and decaf of the sumatra to make a great cup of coffee the coffee has a very rich and bold flavor
5. these candies are not your typical kids sour candies they have exotic flavors and not as tart as the others.

7.Train-Test Split

Train-Test Split is a technique to divide the dataset into two separate sets:

- Training Set (typically 80%): Used to train the machine learning model.

- Test Set (typically 20%): Used to evaluate the model's performance on unseen data.

This separation ensures that the model generalizes well and is not just memorizing the training data.

Why Stratified Split?

In classification problems like review rating prediction, stratified splitting is important to maintain class balance (equal distribution of ratings) in both training and testing sets.

Without stratification, some rating classes (like 1-star or 5-star) may be underrepresented in the test set, leading to biased evaluation.

How it was done

To prepare the data for model training, the dataset was first shuffled randomly to eliminate any order bias. A stratified train-test split was then performed using `train_test_split()` from `sklearn.model_selection` with the `stratify=y` argument to ensure that all star ratings were proportionally represented in both training and testing sets. The dataset was split using an 80% training and 20% testing ratio. After splitting, all text preprocessing steps-including lowercasing, lemmatization, stopwords removal, and cleaning were applied separately to `x_train` and `x_test` to prevent data leakage and maintain model integrity.

Code:

```
from sklearn.model_selection import
train_test_split X =
imbalanced_df["Cleaned_Review"] y =
imbalanced_df["Rating"]
X_train, X_test, y_train, y_test = train_test_split( X, y, test_size=0.2,
random_state=42,
stratify=
```

)

8.Text Vectorization

Machine learning models can't work directly with raw text-they require numerical input. Vectorization is the process of converting text data into numerical features.

TF-IDF (Term Frequency – Inverse Document Frequency)

TF-IDF is a statistical technique that represents how important a word is to a document relative to the entire corpus.

- TF (Term Frequency): How often a word appears in a document.
- IDF (Inverse Document Frequency): Penalizes common words and highlights rare but Important ones.

This approach helps to capture both word relevance and discriminative power, making it better than simple word counts.

Formula for TF-IDF:

TF(t,d)=Total number of terms in document d
Number of times term t appears in

$$TF(t, d) = \frac{\text{number of times term } t \text{ appears in document } d}{\text{total number of terms in document } d}$$

$$IDF(t) = \log \frac{N}{1+df}$$

$$TF - IDF(t, d) = TF(t, d) * IDF(t)$$

Code

```
from sklearn.feature_extraction.text import
TfidfVectorizer import pandas as pd documents = [
    "Machine learning models require numerical input",
    "Text data must be converted into numerical features",
    "TF IDF is a text vectorization technique"
]
vectorizer = TfidfVectorizer()
tfidf_matrix = vectorizer.fit_transform(documents)
feature_names = vectorizer.get_feature_names_out() df =
pd.DataFrame(tfidf_matrix.toarray(), columns=feature_names)
print("TF-IDF feature names (words):") print(feature_names)
print("\nTF-IDF Matrix:") print(df)
```