AUTOMATED REVIEW SYSTEM

1. Project Overview

An intelligent system that analyzes customer reviews and predicts corresponding star ratings. It uses natural language processing (NLP) to extract sentiment and key features from textual feedback. The model is trained on labeled review data to learn patterns between language and rating.

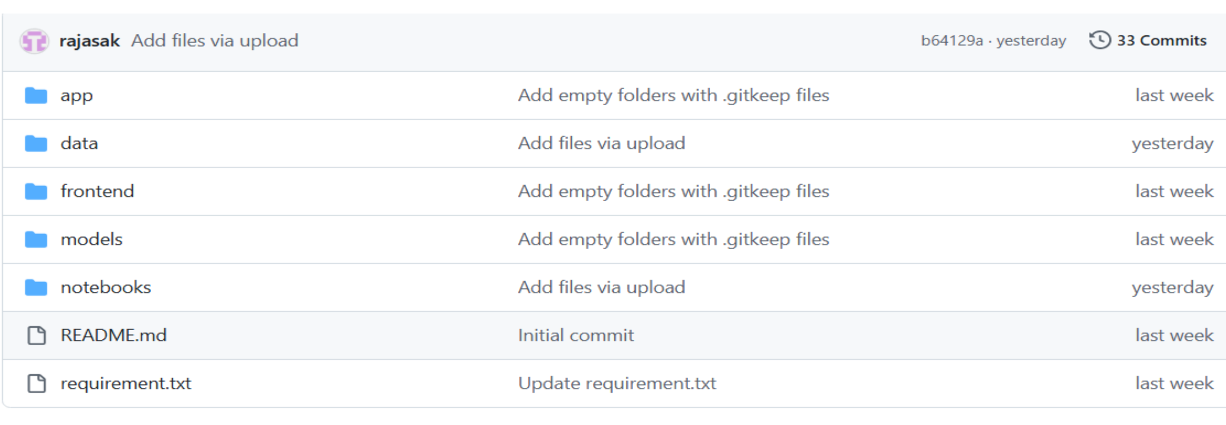
2. Environment Setup

- Installed Python 3.10+ with required libraries: pandas, numpy, matplotlib, seaborn, spacy, scikit-learn, beautifulsoup4, requests
- IDE used: VS Code
- Verified proper functioning of packages for data preprocessing, NLP, and visualization

3. GitHub Project Setup

Created GitHub repository: **automated-review-rating-system**

Structure of directory



4. Data Collection

- Data was collected through web scrapping and from Kaggle
- Data collected from Kaggle had 49000 rows and it was dataset related to cloths
- Data collected from web scrapping had 1700 rows and it was dataset related to iphone 15
- Both the dataset were merged
- Dataset link= https://github.com/rajasak/automated-review-rating-system/blob/main/data/merged.csv
- Final dataset contains 2 column with Review and Rating
- 1 star=4196, 2 star= 3447, 3 star= 5695, 4 star=8316 , 5 star= 29294

4.1. Balanced dataset

- Link for balanced dataset= https://github.com/rajasak/automated-review-rating-system/blob/main/data/balanced_data.csv
- Balanced dataset was created with 2000 rows of each rating

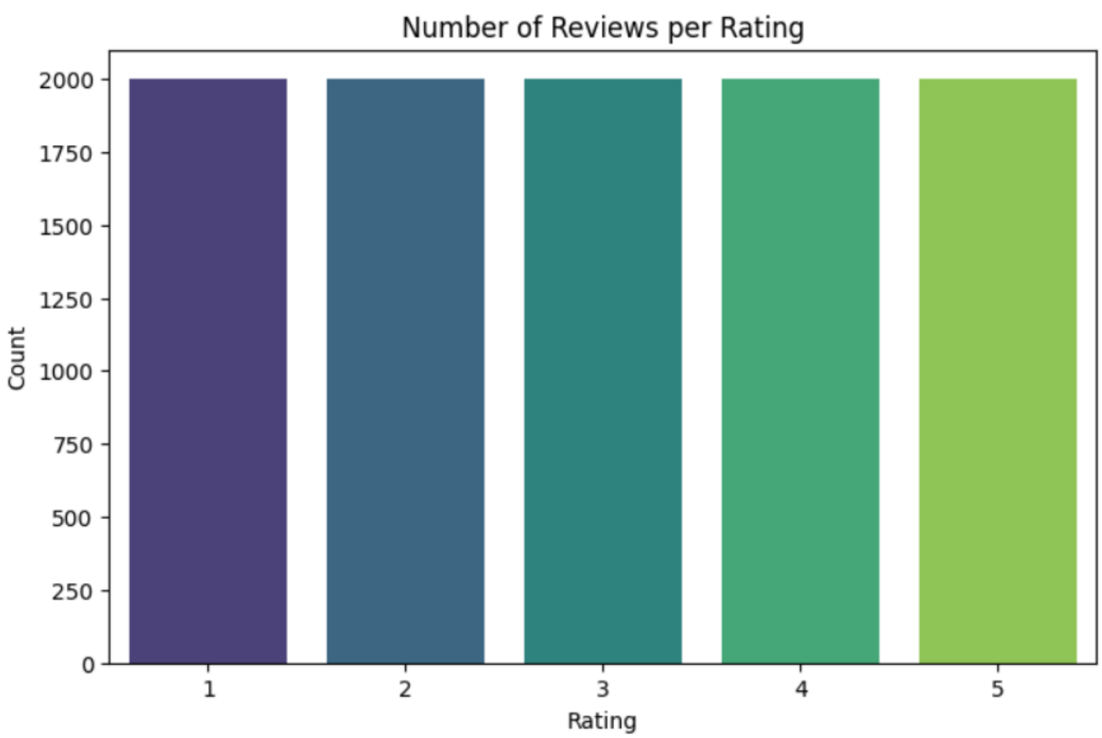


Fig count plot

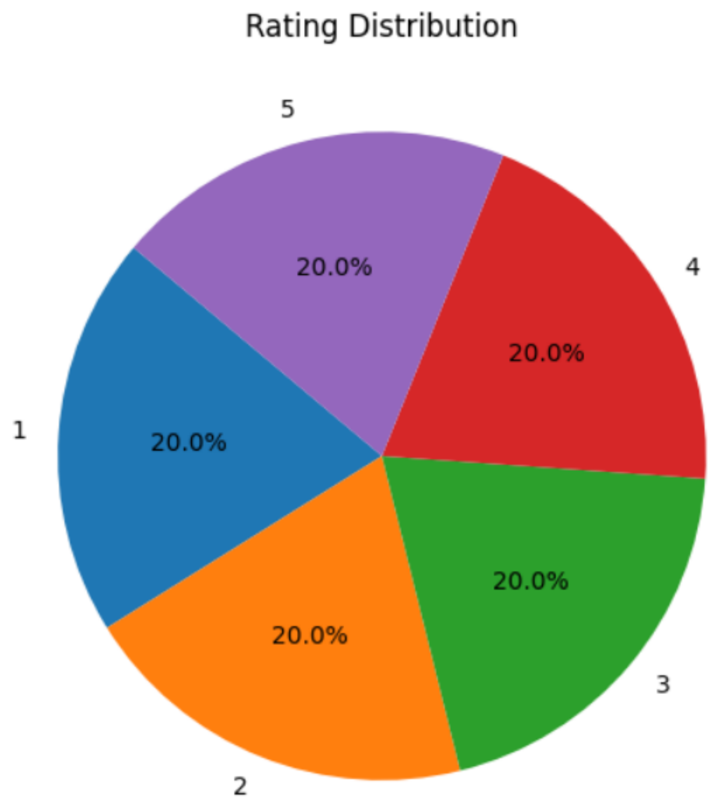


Fig pie chart

4.2. imbalanced dataset

- Link for imbalanced dataset= https://github.com/rajasak/automated-review-rating-system/blob/main/data/imbalanced_ratio_dataset3.csv
- Imbalanced dataset was created with 1star=10%, 2 star=15%, 3 star=15%, 4 star=30%, 5 star=20%

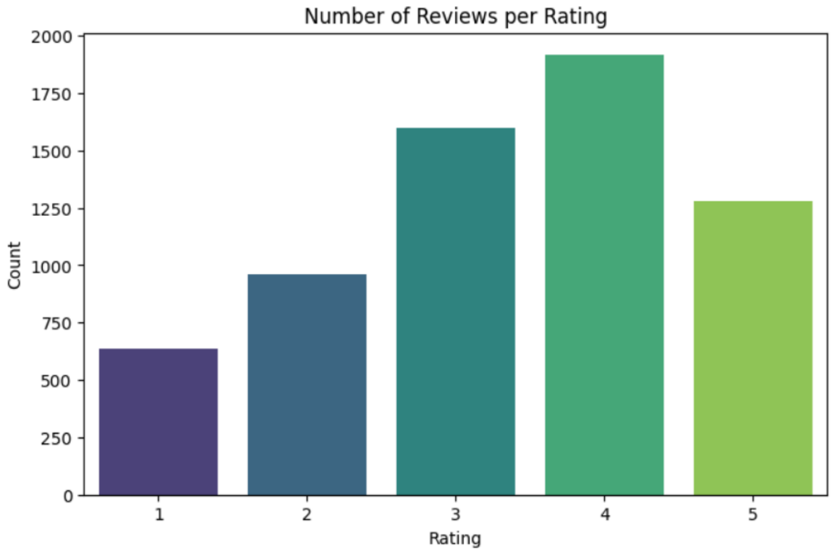


Fig count plot

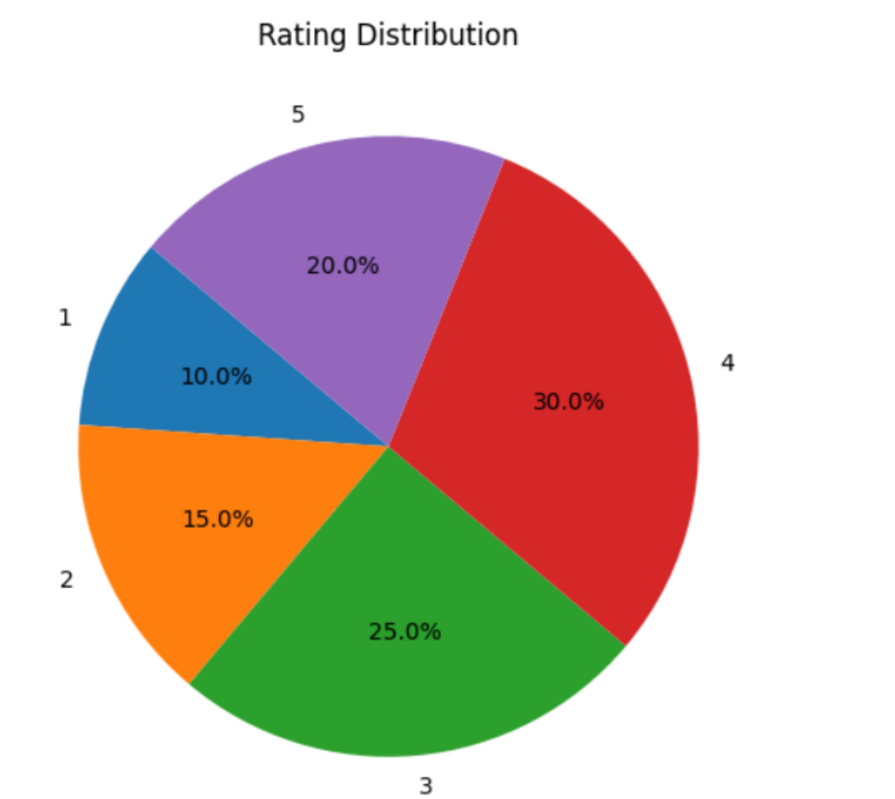


Fig pie chart

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 review 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 observations.

Code:

```
has_duplicates = df.duplicated().any()
print(has_duplicates)

num_duplicates = df.duplicated().sum()
print(f"Number of duplicate rows: {num_duplicates}")

df.drop_duplicates(inplace=True)
```

5.2 Removing Conflicting Reviews

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

code:

```
# Filter to only reviews that have more than one unique rating
conflicting_reviews = conflicting_counts[conflicting_counts > 1]
print("Number of conflicting review texts:", conflicting_reviews.shape[0])


conflicting_reviews = df.groupby('Review')['Rating'].nunique()
conflicting_reviews = conflicting_reviews[conflicting_reviews > 1].index
df = df[df['Review'].isin(conflicting_reviews) == False]
```

5.3 Handling Missing Values

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

Code:

```
df = df.dropna(subset=['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 Lowercasing Text

All review text was converted to lowercase to maintain uniformity. This helps prevent duplication of tokens like "Good" and "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

Emojis and special symbols may not contribute meaningful context for text analysis (unless specifically required). We remove these characters to focus on the core textual content.

5.8 Removing Punctuation

Punctuation marks and numbers can disrupt the natural language patterns of the text. By removing them, we simplify the tokenization process and prevent punctuation from being misinterpreted as separate tokens.

Code:

```
#removing punctuations and making lower case
X_train = X_train.str.lower()
X_train = X_train.str.replace(r'\[.*?\]', '', regex=True)


import re

def clean_text(text):
    text = str(text)
    text = re.sub(r"http\S+|www\S+|https\S+", '', text, flags=re.MULTILINE)
# remove URLs
    text = re.sub(r'<.*?>', '', text)  # remove HTML tags
    text = re.sub(r'[\U00010000-\U0010ffff]', '', text)  # remove emojis
    text = re.sub(r'[^\w\s]', '', text)  # remove punctuation
    text = re.sub(r'[^a-zA-Z0-9\s]', '', text)  # remove special characters
    text = re.sub(r'\s+', ' ', text).strip()  # remove extra whitespace
    return text
```

5.9 Stopwords Removal

Stopwords, such as "the", "is", and "and", are frequently occurring words that might not add significant semantic value. We use libraries like SpaCy to filter these words out, reducing noise and focusing on words that contribute meaning to the reviews, there were total 326 stop words in spacy.

```
["'d", "'ll", "'m", "'re", "'s", "'ve", 'a', 'about', 'above', 'across', 'after', 'afterwards', '
again', 'against', 'all', 'almost', 'alone', 'along', 'already', 'also', 'although', 'always', 'a
m', 'among', 'amongst', 'amount', 'an', 'and', 'another', 'any', 'anyhow', 'anyone', 'anything',
'anyway', 'anywhere', 'are', 'around', 'as', 'at', 'back', 'be', 'became', 'because', 'become', '
becomes', 'becoming', 'been', 'before', 'beforehand', 'behind', 'being', 'below', 'beside', 'besi
des', 'between', 'beyond', 'both', 'bottom', 'but', 'by', 'ca', 'call', 'can', 'cannot', 'could',
'did', 'do', 'does', 'doing', 'done', 'down', 'due', 'during', 'each', 'eight', 'either', 'eleven
', 'else', 'elsewhere', 'empty', 'enough', 'even', 'ever', 'every', 'everyone', 'everything', 'ev
erywhere', 'except', 'few', 'fifteen', 'fifty', 'first', 'five', 'for', 'former', 'formerly', 'fo
rty', 'four', 'from', 'front', 'full', 'further', 'get', 'give', 'go', 'had', 'has', 'have', 'he'
, 'hence', 'her', 'here', 'hereafter', 'hereby', 'herein', 'hereupon', 'hers', 'herself', 'him',
'himself', 'his', 'how', 'however', 'hundred', 'i', 'if', 'in', 'indeed', 'into', 'is', 'it', 'it
s', 'itself', 'just', 'keep', 'last', 'latter', 'latterly', 'least', 'less', 'made', 'make', 'man
y', 'may', 'me', 'meanwhile', 'might', 'mine', 'more', 'moreover', 'most', 'mostly', 'move', 'muc
h', 'must', 'my', 'myself', "n't", 'name', 'namely', 'neither', 'never', 'nevertheless', 'next',
'nine', 'no', 'nobody', 'none', 'noone', 'nor', 'not', 'nothing', 'now', 'nowhere', 'n‘t', 'n’t',
'of', 'off', 'often', 'on', 'once', 'one', 'only', 'onto', 'or', 'other', 'others', 'otherwise',
'our', 'ours', 'ourselves', 'out', 'over', 'own', 'part', 'per', 'perhaps', 'please', 'put', 'qui
te', 'rather', 're', 'really', 'regarding', 'same', 'say', 'see', 'seem', 'seemed', 'seeming', 's
eems', 'serious', 'several', 'she', 'should', 'show', 'side', 'since', 'six', 'sixty', 'so', 'som
e', 'somehow', 'someone', 'something', 'sometime', 'sometimes', 'somewhere', 'still', 'such', 'ta
ke', 'ten', 'than', 'that', 'the', 'their', 'them', 'themselves', 'then', 'thence', 'there', 'the
reafter', 'thereby', 'therefore', 'therein', 'thereupon', 'these', 'they', 'third', 'this', 'thos
e', 'though', 'three', 'through', 'throughout', 'thru', 'thus', 'to', 'together', 'too', 'top', '
toward', 'towards', 'twelve', 'twenty', 'two', 'under', 'unless', 'until', 'up', 'upon', 'us', 'u
sed', 'using', 'various', 'very', 'via', 'was', 'we', 'well', 'were', 'what', 'whatever', 'when',
'whence', 'whenever', 'where', 'whereafter', 'whereas', 'whereby', 'wherein', 'whereupon', 'where
ver', 'whether', 'which', 'while', 'whither', 'who', 'whoever', 'whole', 'whom', 'whose', 'why',
'will', 'with', 'within', 'without', 'would', 'yet', 'you', 'your', 'yours', 'yourself', 'yoursel
ves', '‘d', '‘ll', '‘m', '‘re', '‘s', '‘ve', '’d', '’ll', '’m', '’re', '’s', '’ve']
```

Code:

```
#lemmatization and stop word removal
def cleaning(text):
    doc = nlp(text)
    return ' '.join([token.lemma_ for token in doc if not token.is_stop])
```

5.10 Lemmatization

Lemmatization is a text preprocessing technique that reduces words to their base or dictionary form, known as the **lemma**. For example, words like **"running"**, **"ran"**, and **"runs"** are all reduced to the base form **"run"**. Unlike stemming, lemmatization uses linguistic rules and vocabulary, ensuring that the resulting word is meaningful. This helps in grouping similar words and improves model performance by reducing the size of the vocabulary without losing context.

Why lemmatization is better than stemming

Lemmatization is often preferred over stemming because:

1. **Produces Real Words**:

   Lemmatization returns valid dictionary words (e.g., *"running"* → *"run"*), whereas stemming may return incomplete or non-existent words (e.g., *"running"* → *"runn"*).

2. **Context-Aware**:

   Lemmatization considers the context and part of speech (POS) of a word to find its correct root form.
   Stemming blindly trims word endings without understanding grammar.

3. **More Accurate and Clean**:

   Lemmatization provides cleaner and more accurate tokens, which improves model understanding and performance, especially in NLP tasks like sentiment analysis or classification.

5.11 Filtering by Word Count

Very short reviews might not contain enough context to be useful, and excessively long reviews could be outliers. We apply filtering to exclude:

- Reviews with fewer than 3 words.
- Reviews that exceed 100 words.
  This ensures that the dataset remains robust and relevant for model training.

6. Data visualization

Box plot

A **box plot** is a visual tool used to display the distribution of numerical data, showing the median, quartiles, and potential outliers. It helps quickly understand how data is spread and identify any extreme values. In this project, box plots were used to compare the word count distribution of reviews across different rating classes.

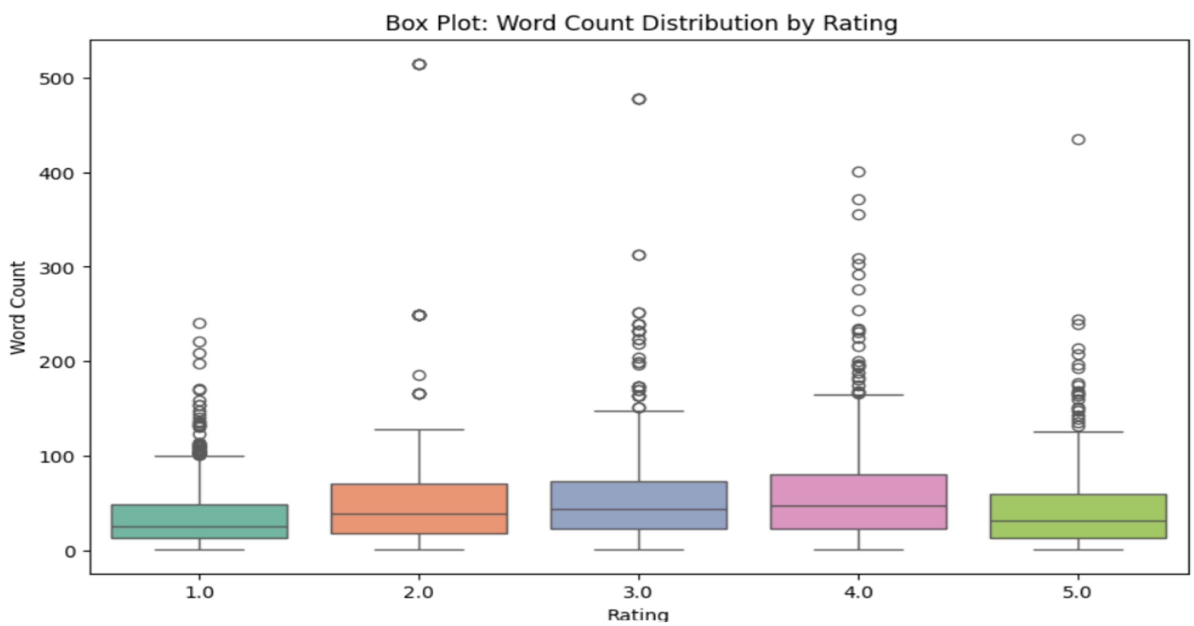


Fig box plo

Histogram

A **histogram** is a graphical representation that shows the distribution of a numeric variable by dividing the data into intervals (bins) and counting how many values fall into each bin. In this project, histograms were used to visualize the **word count distribution** of reviews, helping to identify common review lengths and spot patterns across different rating levels.

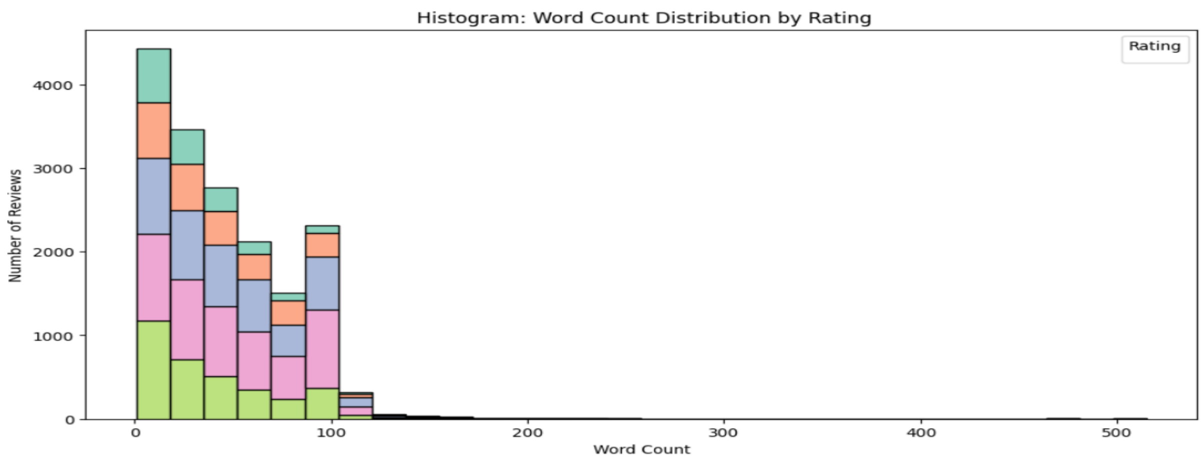


Fig histogram

Barplot

A **bar plot** is a chart that represents categorical data with rectangular bars, where the height of each bar indicates the value or frequency of that category. In this project, bar plots were used to show the number of reviews per rating class (1 to 5 stars)**,** helping to visualize class distribution and detect any imbalance in the dataset.

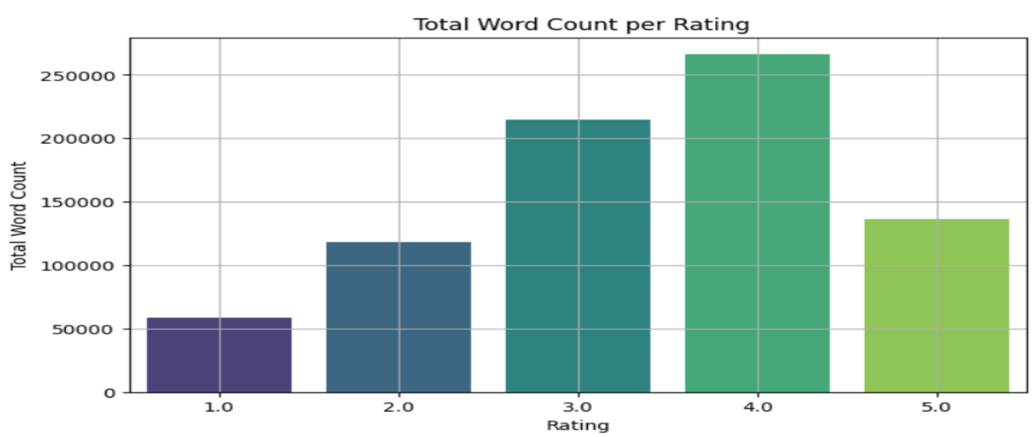


Fig bar plot

Samples of 1 star rating:

Showing 5 sample reviews for Rating 1:

1. I wore these pants to work and looked great! however the first time i put them through the wash they became all discolored and are completely stained and now are unwearable. very disappointing. did not spend that much money to wear a pair of pants once.

2. This sweater is enormous. I expected it to be oversized, but I ordered a size small (5’6, 125lbs) and it fit like an XL. The sweat er also had a very bad smell to it. Returning.

3. This thing runs extremely hot, is massively oversized, and doesn’t feel soft and silky on the inside at all. Returning it.

4. Absolutely no stretch, very stiff and thick material, way too short, and shrinks quickly. Selling this item on Poshmark if anyone is interested at a discounted price @homeandcloset.

5. Retailer consistently provides unique, sophisticated and quality made clothing. with the exception of an occasional purchase at their sister companies, i shop with them exclusively. i suppose that i can't expect 100% perfection from retailer-but 99% of the ti me-that's what i get. this is that 1%...this sweater is truly awful! i cannot (or find it very difficult at least) begin to describe what t his sweater is like but it is so unlike retailer that, had it not been for the distinct pattern on the front

Samples of 2 star rating

Showing 5 sample reviews for Rating 2:

1. The color was not white it was off white and the fabric was very flimsy

2. This is such a beautiful dress. but have to return for the reasons others have noted. the chest does not fit well.

3. not perma press!has to be ironed!

4. I like the design of the dress, but the fabric makes it look cheap. for an expensive dress, i had expected better quality.

5. Must return. It looked just right on the model and an embarrassment on me. Even after wash/dry on delicate and immediate dry er removal it was too wrinkled to wear without ironing. Reread the fabric description on Amazon...says "polyester and spandex", but the tag on the blouse includes RAYON. This would have to be ironed before wearing. Sooo disappointing.

Samples of 3 star rating

Showing 5 sample reviews for Rating 3:

1. Couldn't wait for this sweater to arrive. i love oversized sweaters for the fall! unfortunately this sweater is extremely itchy that it has to be returned :(

2. As much as i had wanted this to work for me, it didn't. fabric is not as soft as i had expected and quality is just bad. i purchased this online (on sale at $29.95) and i had envisioned this to be comfy stretchy and swingy with a flow. i sent this back to the store as it didn't seem fit even at the low price.

3. I wore it. Probably will find its way into a give-away bag. Just too swingy; too much fabric. Not for the short gals; not for the busty gals; not for the self-conscious "hope no one is staring at my middle" gals.

4. Ordered medium as usual, way small

5. I first have to agree with other reviewers about the quality of this sweater: the stitch is pretty, but the fabric isn't nearly as soft a s i expected. it has the look and feel of a cheaper sweater. that said, the fit is pretty flattering. i didn't have any problem getting th e cowl over my shoulders as pictured by the model. i've been loo

Samples of 4 star rating

Showing 5 sample reviews for Rating 4:

1. I was looking for some comfortable but still stylish shorts for the weekend. these hit the mark!

2. Las medidas se ajustan perfectamente con la realidad

3. Ordered both the black and white and ordered both in 0 and p2. preferred the black over white and the p2 fit the best. however, they reminded me too much of a maternity top so they both went back.
if you order you will need to wear a cami underneath?at least i would. not a terrible top, just not for me.

4. Just note the pink is different than in the photos. much hotter/more saturated. apart from that surprise, overall i'd recommend. n ice design.

5. Like another reviewer said this was too baggy up top. the bottom length was good but arm holes were big and top was just too l arge. i'm 5'5" 110 lbs and got xs. returned.

Samples of 5 star rating

Showing 5 sample reviews for Rating 5:

1. Fit like regular wranglers but with some added stretch in the waits for those thanksgiving dinners!

2. I saw this duster and immediately ordered it in cream. i have worn it every chance i get. it goes well with dresses as well as shorts or jeans. the length is perfect, the design is flattering, and the material is soft and comfortable. the sleeves are snug- but i like this feature, especially for light weight summer use over a tank top or sleeveless blouse. this will be a go to item. thanks retailer- another gem!

3. good fabric greatfit and the price is right

4. Very nice. Material is comfortable.

5. So I knew that the pajama bottoms had squirrels on them, but no idea that those squirrels were wearing scarves, ear muffs, Santa hats and booties! I adore the pattern! I ordered the size L because I like my jammies loose. I’d say that the sizing runs a little small because I thought the top might be a little too big (I’m bigger on the bottom than the top), but it’s not - it fits perfectly.

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, stopword removal, and cleaning—were applied **separately** to `X_train` and `X_test` to **prevent data leakage** and maintain model integrity.

Code:

```python
# Shuffle the balanced DataFrame
df = df.sample(frac=1, random_state=42).reset_index(drop=True)


from sklearn.model_selection import train_test_split


x=df['Review']
y=df['Rating']


X_train, X_test, y_train, y_test = train_test_split(
    x, y, test_size=0.2, random_state=42, stratify=y
)
```

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.

<u>Formula for TF-IDF:</u>

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

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

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

Code:

```
from sklearn.feature_extraction.text import  TfidfVectorizer
# create a tfidf vectorizer matrix
tv = TfidfVectorizer()
Xtrain = tv.fit_transform(X_train)
X = pd.DataFrame(Xtrain.toarray(), columns=tv.get_feature_names_out())
X
```

Output:

Out[165...

	00	000	002first	00p	02	025	04	045blue	045purple	0dd	...	zipper	zippered	zipperi	zipperit	zipperone	zips	zipu
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.0	0.0	0.0	0.0	0.0	0.
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.0	0.0	0.0	0.0	0.0	0.
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.0	0.0	0.0	0.0	0.0	0.
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.0	0.0	0.0	0.0	0.0	0.
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.0	0.0	0.0	0.0	0.0	0.
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	.
13090	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.0	0.0	0.0	0.0	0.0	0.
13091	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.444043	0.0	0.0	0.0	0.0	0.0	0.
13092	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.0	0.0	0.0	0.0	0.0	0.
13093	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.0	0.0	0.0	0.0	0.0	0.
13094	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.0	0.0	0.0	0.0	0.0	0.

9. Training the model

Algorithm: Logistic Regression

Logistic Regression is a supervised learning algorithm used primarily for classification tasks. In this project, it is used for multi-class text classification to predict review ratings (1–5 stars) based on review content. Despite its name, Logistic Regression is a classification, not regression, algorithm.

How the Algorithm Works

Binary Classification

For binary classification, logistic regression models the probability that an input x belongs to class 1:

$$z = \mathbf{w}^T\mathbf{x} + b$$

$$P(y = 1 \mid \mathbf{x}) = \sigma(z) = \frac{1}{1 + e^{-z}}$$

Where:

- w: weight vector
- x: feature vector (e.g., TF-IDF)
- b: bias term
- σ(z): sigmoid function that maps z to (0, 1)

Prediction is made by:

$$\hat{y} = \begin{cases} 1 & \text{if } \sigma(z) \geq 0.5 \\ 0 & \text{otherwise} \end{cases}$$

Multi-Class Classification (Softmax Regression)

When there are more than two classes (e.g., 5 review rating levels), logistic regression uses the **softmax function** to compute the probability of each class k ∈ {1,2,...,K}:

$$z_k = \mathbf{w}_k^T \mathbf{x} + b_k$$

$$P(y = k \mid \mathbf{x}) = \frac{e^{z_k}}{\sum_{j=1}^{K} e^{z_j}} \quad \text{for } k = 1, 2, ..., K$$

The model predicts the class with the **highest probability**:

$$\hat{y} = \arg\max_k \; P(y = k \mid \mathbf{x})$$

Key Hyperparameter:

Hyperparameter	Description	Default	Tuning Tips
C	Inverse of regularization strength	1.0	Lower = stronger regularization
penalty	Type of regularization (`l1`, `l2`, etc.)	l2	Use l2 for text data
solver	Optimization algorithm (liblinear, saga, etc.)	lbfgs	Use saga for l1 or large datasets
multi_class	Strategy: ovr (one-vs-rest) or multinomial (softmax)	auto	Use multinomial with softmax logic
max_iter	Max optimization iterations	100	Increase if not converging

Strengths:

- Fast training and prediction
- Works well with high-dimensional sparse data (like TF-IDF)
- Easy to interpret (coefficients show feature importance)
- Performs well on linearly separable classes

Limitations:

- Assumes linear decision boundary
- Not ideal for complex, nonlinear relationships
- Sensitive to irrelevant features and multicollinearity
- Performance drops with highly imbalanced classes

Use When:

- Dataset is large and well-preprocessed
- You need a fast, interpretable baseline
- Input features are TF-IDF or BoW vectors

Avoid When:

- Data is highly nonlinear or hierarchical
- High accuracy is required for long-term deployment
- You want contextual embeddings (BERT, LSTM, etc.)

Model Evaluation: Logistic Regression Results:

Code:

```
# Train Logistic Regression
lr = LogisticRegression(max_iter=1000)
lr.fit(Xtrain_vec, y_train)

# Predict
y_pred_lr = lr.predict(Xtest_vec)

# Evaluation
print(" Logistic Regression:")
print("Accuracy:", accuracy_score(y_test, y_pred_lr))
print("Classification Report:\n", classification_report(y_test, y_pred_lr))
```

Result:

```
 Logistic Regression:
Accuracy: 0.550561797752809
Classification Report:
              precision    recall  f1-score   support

         1.0       0.61      0.38      0.47       323
         2.0       0.62      0.48      0.54       490
         3.0       0.53      0.62      0.57       832
         4.0       0.50      0.58      0.54       994
         5.0       0.62      0.55      0.58       654

    accuracy                           0.55      3293
   macro avg       0.58      0.52      0.54      3293
weighted avg       0.56      0.55      0.55      3293
```

After training the Logistic Regression model on the text data using TF-IDF features, we evaluated its performance on the test set using standard classification metrics.

- **Accuracy:** 0.550 (55%)
  This indicates that the model correctly predicted the review class in about 55% of the test samples, which is a reasonable baseline for a 5-class classification task.
- **Macro Average F1-Score:** 0.54
  A simple average over all classes, treating each class equally.
- **Weighted Average F1-Score:** 0.55
  Accounts for the support (number of instances) of each class, giving a more balanced overall view.
- The model performs **best on class 5.0**, with an F1-score of **0.58**, indicating it is best at identifying highly positive reviews.
- Classes **1.0** have the **lowest performance**, which suggests overlap in feature space or ambiguity in the reviews that correspond to neutral sentiments.