Assignment2 Report

Diya Khurdiya, 1020201649

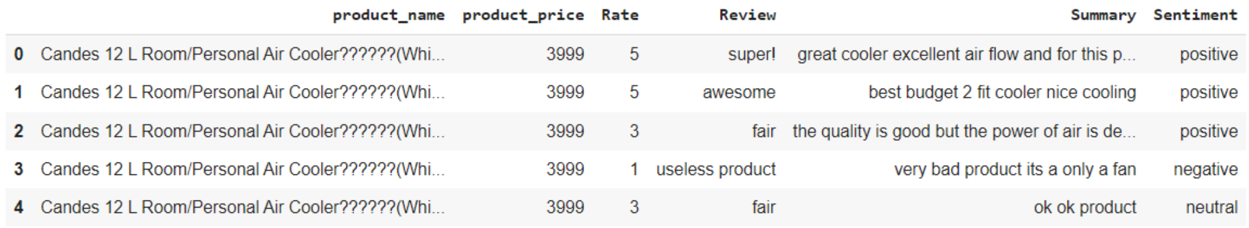
Sentiment Analysis

Introduction:

Sentiment analysis falls under the heading of text classification and is a use case of natural language processing (NLP). Simply described, sentiment analysis includes categorising a text into several emotions, such as happy or sad, neutral, or happy or sad. Determining the underlying tone, emotion, or sentiment of a document is the ultimate goal of sentiment analysis.

- The Flipkart dataset provides us features like product_price, product_name, reviews and summaries about the products.
- It has 205053 rows and 6 columns.
- Product_name: Name of the product.
  Product_price:Price of the product.
  Rate: Customer's rating on product(Between 1 to 5).
  Review: Customer's review on each product.
  Summary: This columne include descriptive informationn of customer's thought on each product.
  Sentiment: This column contains 3 label such as Positive, Negative and Neutral(Which was given based on Summary).

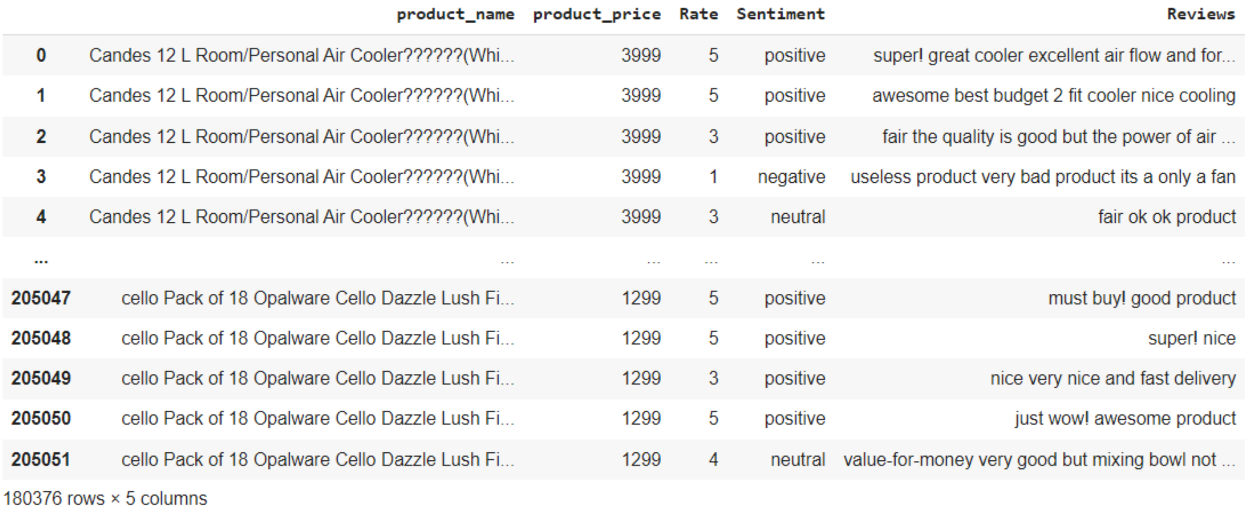
Glimpse of the dataset



	product_name	product_price	Rate	Review	Summary	Sentiment
0	Candes 12 L Room/Personal Air Cooler??????(Whi...	3999	5	super!	great cooler excellent air flow and for this p...	positive
1	Candes 12 L Room/Personal Air Cooler??????(Whi...	3999	5	awesome	best budget 2 fit cooler nice cooling	positive
2	Candes 12 L Room/Personal Air Cooler??????(Whi...	3999	3	fair	the quality is good but the power of air is de...	positive
3	Candes 12 L Room/Personal Air Cooler??????(Whi...	3999	1	useless product	very bad product its a only a fan	negative
4	Candes 12 L Room/Personal Air Cooler??????(Whi...	3999	3	fair	ok ok product	neutral

Pre-processing:

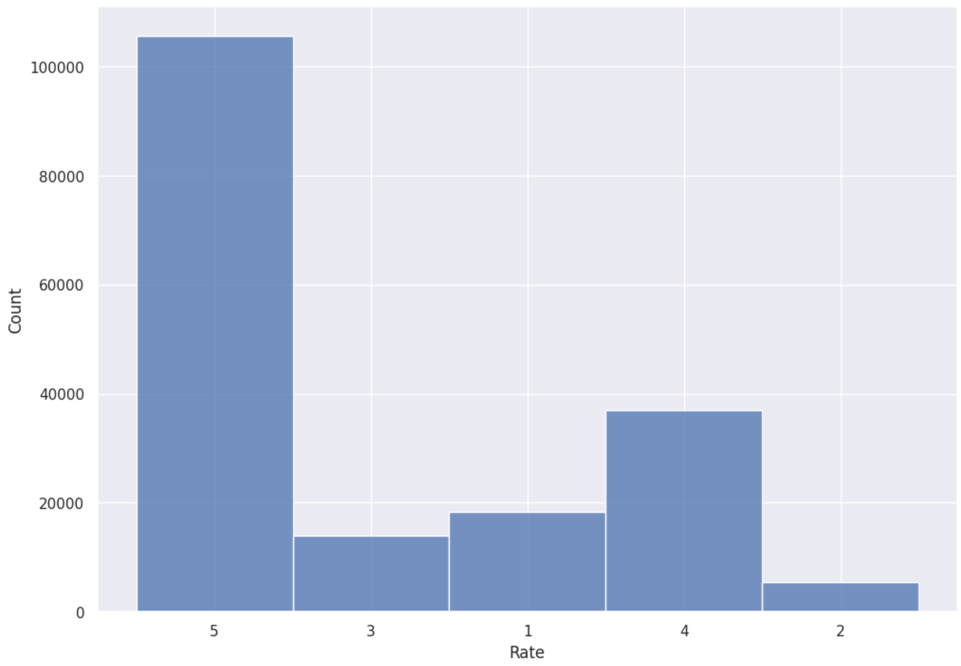
The data needs pre-processing as there are some null values and there is irrelevant text.
We also concatenate review and summary so as to form a common thread for evaluation. Garbage text in the review and price section had to be removed.



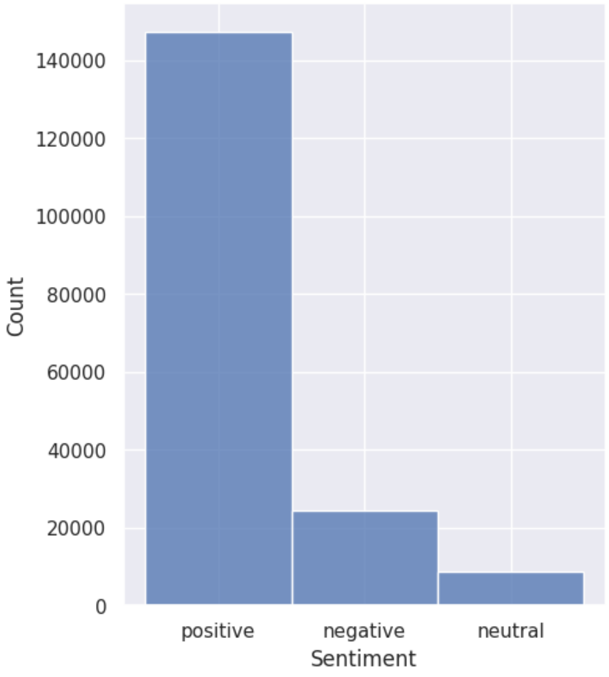
	product_name	product_price	Rate	Sentiment	Reviews
0	Candes 12 L Room/Personal Air Cooler??????(Whi...	3999	5	positive	super! great cooler excellent air flow and for...
1	Candes 12 L Room/Personal Air Cooler??????(Whi...	3999	5	positive	awesome best budget 2 fit cooler nice cooling
2	Candes 12 L Room/Personal Air Cooler??????(Whi...	3999	3	positive	fair the quality is good but the power of air ...
3	Candes 12 L Room/Personal Air Cooler??????(Whi...	3999	1	negative	useless product very bad product its a only a fan
4	Candes 12 L Room/Personal Air Cooler??????(Whi...	3999	3	neutral	fair ok ok product
...	...	...	...	...	...
205047	cello Pack of 18 Opalware Cello Dazzle Lush Fi...	1299	5	positive	must buy! good product
205048	cello Pack of 18 Opalware Cello Dazzle Lush Fi...	1299	5	positive	super! nice
205049	cello Pack of 18 Opalware Cello Dazzle Lush Fi...	1299	3	positive	nice very nice and fast delivery
205050	cello Pack of 18 Opalware Cello Dazzle Lush Fi...	1299	5	positive	just wow! awesome product
205051	cello Pack of 18 Opalware Cello Dazzle Lush Fi...	1299	4	neutral	value-for-money very good but mixing bowl not ...

180376 rows × 5 columns

After cleaning null values, and having reviews and summary under one heading, we proceed for some basic analyze to get a sense of the data given.



The frequency bar chart shows us that maximum reviews have been rated as 5 star.



Similarly, frequency bar plot for sentiments also makes sense as the highest number of reviews that have got 5 star rating are likely to be positive responses.

Now, to have a more accurate analysis and classification we further break down the review keywords by,

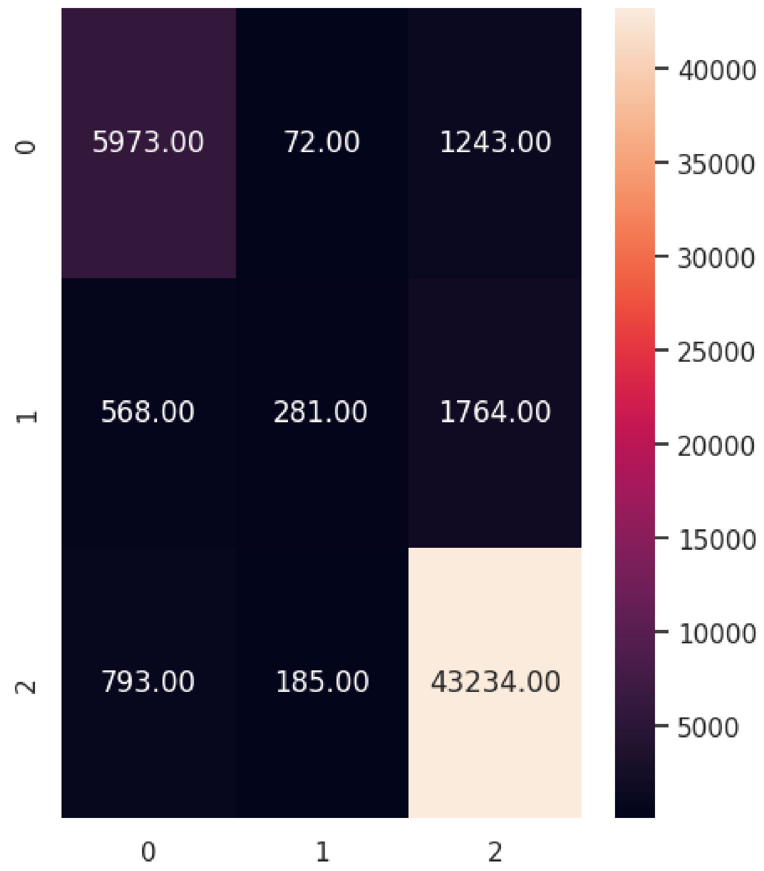
- Tokenization: Splitting the words of a sentence into tokens
- Lemmatization: Converting the word to root words for similar identities
- Stopword removal: Removing words of irrelevance like ‘a’, ‘the’, ‘to’ etc
- Punctuation removal: Removing punctuation marks and making the sentence case insensitive

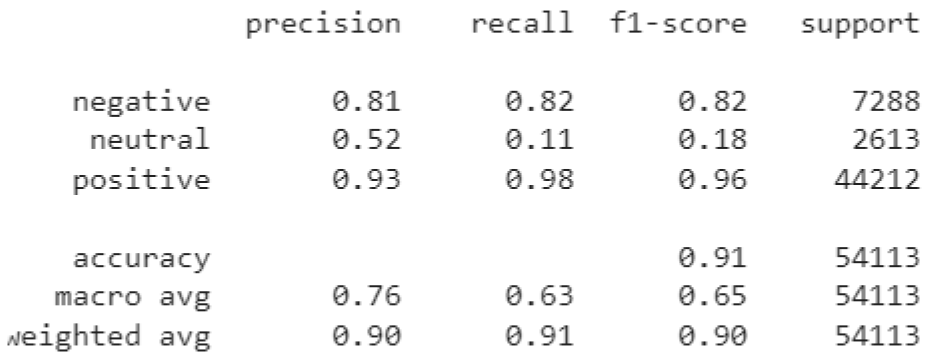
Training and testing:

We now split the dataset as 70% for training and 30% for testing using sklearn train_test_split function.
For our machine learning classification we will be using Multinomial Naive Bayes as that works along the principle of Bayes theorem and is suitable to classify with discrete features. It’s much easier in terms of implementation and can handle large datasets with ease.

We then use Count_Vectorizer to vectorize the text i.e. transforms a token into a vector. We create a sparse array matrix of numeric values and now any machine learning algorithm can be used to sort of train the data.

We then plot the results and confusion matrix, metric report as follows:





```
              precision    recall  f1-score   support

    negative       0.81      0.82      0.82      7288
     neutral       0.52      0.11      0.18      2613
    positive       0.93      0.98      0.96     44212

    accuracy                           0.91     54113
   macro avg       0.76      0.63      0.65     54113
weighted avg       0.90      0.91      0.90     54113
```

We have achieved an accuracy of 90%.

As seen from the SentimentAnalyzer polariser score, positive reviews generally have a high positive score than negative or neutral.

We have a dictionary or bag of models that has pre-classified set of positive and negative keywords. Now, after cleaning each sentence

identify the keywords and assign a score of +1 for positive, -1 for negative and 0 for neutral.
For example, if we have the comment: ‘The product is great’.
After text pre-processing and cleaning we get ‘product great’, now product gets the score 0 for neutral and great gets +1. Hence, the total score becomes 0 + 1 = +1 (positive)

In this case, positive sentiment has been assigned to words like -

	features	counts
1227	wonderful	32043
494	good	16158
110	awesome	11073
1105	terrific	10773
767	nice	8652
1048	specified	7850
918	recommended	6516
500	great	5607