Sentiment Analysis Task Report

Overview

The objective of the assignment is to develop an approach to identify subthemes and their respective sentiments from given reviews. The approach needs to extract insights from the reviews, identifying specific aspects (subthemes) and their associated sentiments, and then visualize these results.

Approach

The approach involves multiple steps, from data preprocessing and sentiment analysis to visualization and topic modeling. Each part of the provided code contributes to achieving the objective.

1. Data Loading/Preprocessing - Start the code by loading the dataset with `pd.read_csv`. A column is removed for cleaning. The data distribution is shown by counting each column's unique values.

2. Sentiment Analysis - Count positive and negative attitudes across all columns. Iterate through each row and check for 'positive' and 'negative' keywords.

3. Visualization - Data is visualized through plots:

- A pie chart reflecting positive evaluations.
- A heatmap showing good reviews in the top 10 categories.
- A line plot of positive review patterns across columns.
- A bar chart showing positive and negative mood.
- Pie and bar charts showing sentiment across rows.

4. Data Cleaning - Lowercase text data in each column, removing punctuation and digits.

5. Subtheme Identification using Pretrained Models VADAR, Roberta and Topic Modeling (LDA)

- Performing sentiment analysis using the VADER model and comparing it with the RoBERTa model helps in quantifying sentiments in the dataset. This step is crucial for determining the sentiment polarity towards each subtheme identified from the reviews.

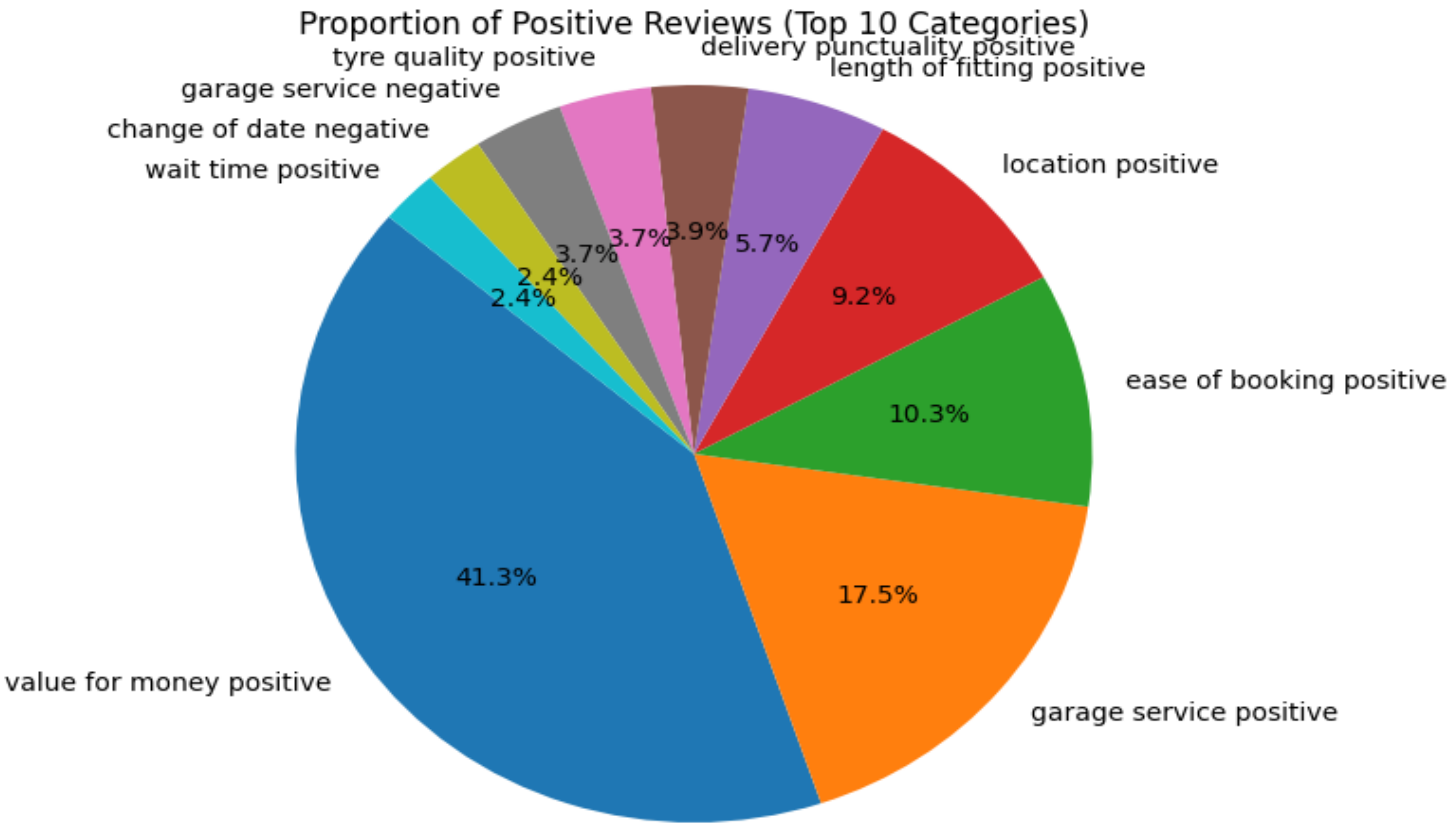
Topic Modelling with LDA:

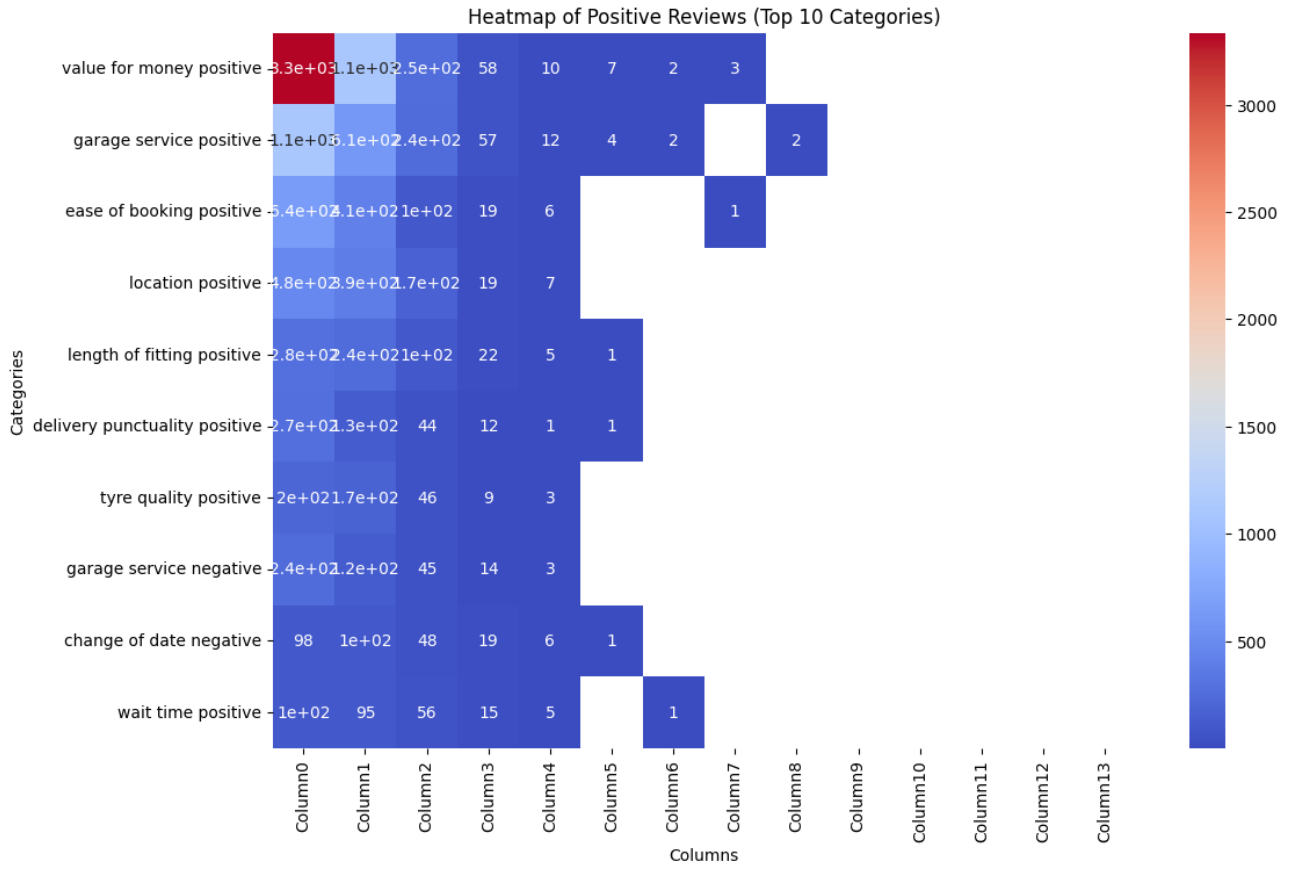
- Utilizing Latent Dirichlet Allocation (LDA) for topic modeling helps in identifying key topics or subthemes present in the text data. This is directly relevant to the task of identifying subthemes from reviews along with their respective sentiments.

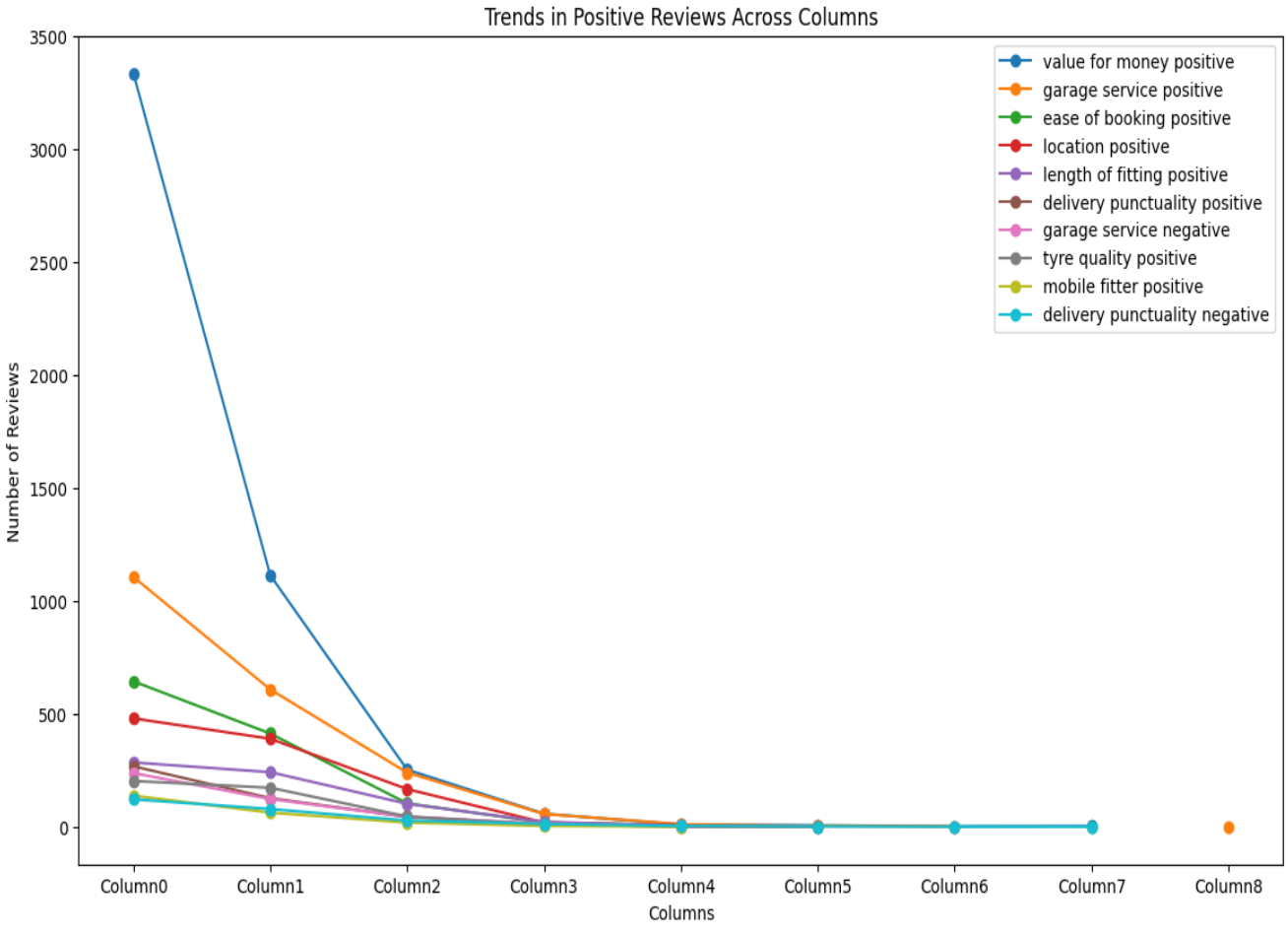
Conclusion:

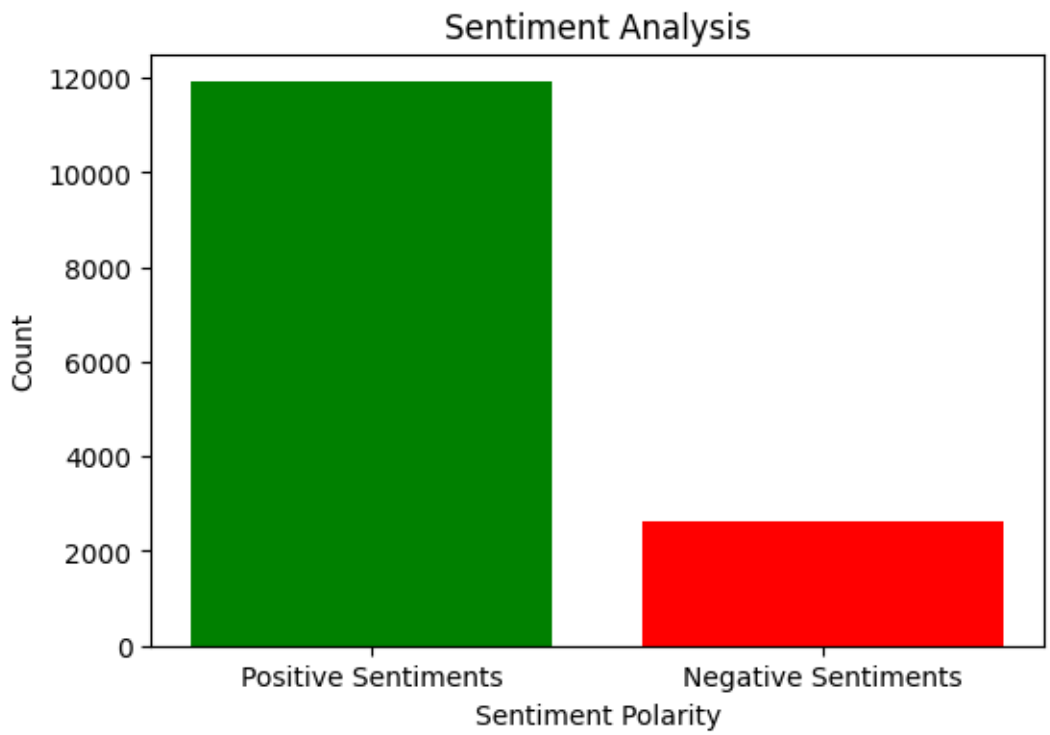
Results include:

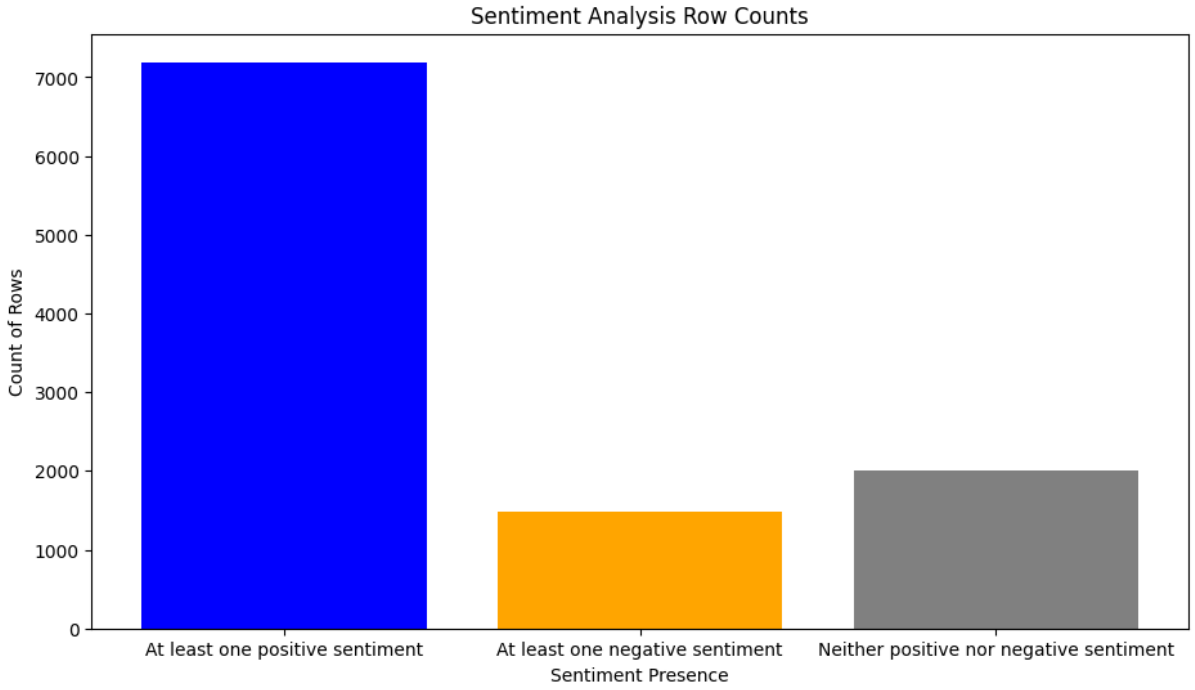
- Positive and negative sentiments are counted.
- Row counts by sentiment presence: The number of rows with one positive, one negative, or neither sentiment is calculated.
- Sentiment distributions and trends are visualized using multiple charts.
- Subtheme Classification: The RoBERTa model may classify subthemes using sentiment ratings, though not explicitly.
- LDA: Perform topic modeling to identify subthemes in the text data.
- Extracts key topics from the reviews, aiding in the identification of subthemes.

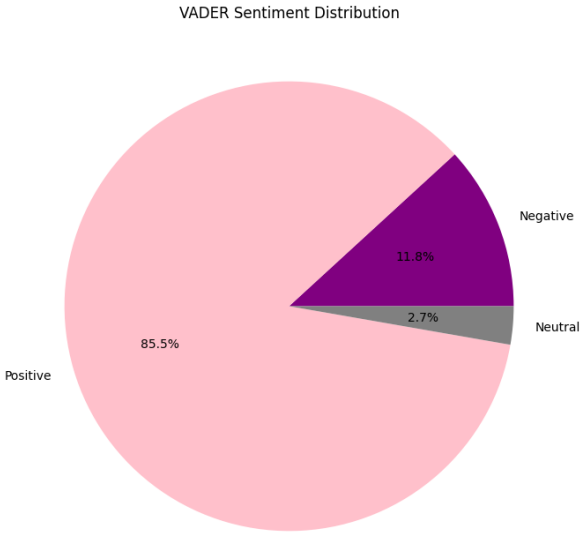


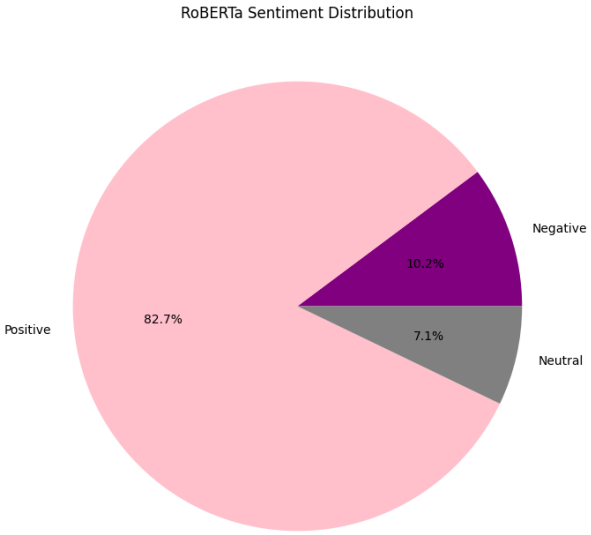

Heatmap of Positive Reviews (Top 10 Categories)
value for money positive
garage service positive
ease of booking positive
location positive
length of fitting positive
delivery punctuality positive
tyre quality positive
garage service negative
change of date negative
wait time positive
Categories
Column0
Column1
Column2
Column3
Column4
Column5
Column6
Column7
Column8
Column9
Column10
Column11
Column12
Column13
Columns
3000
2500
2000
1500
1000
500

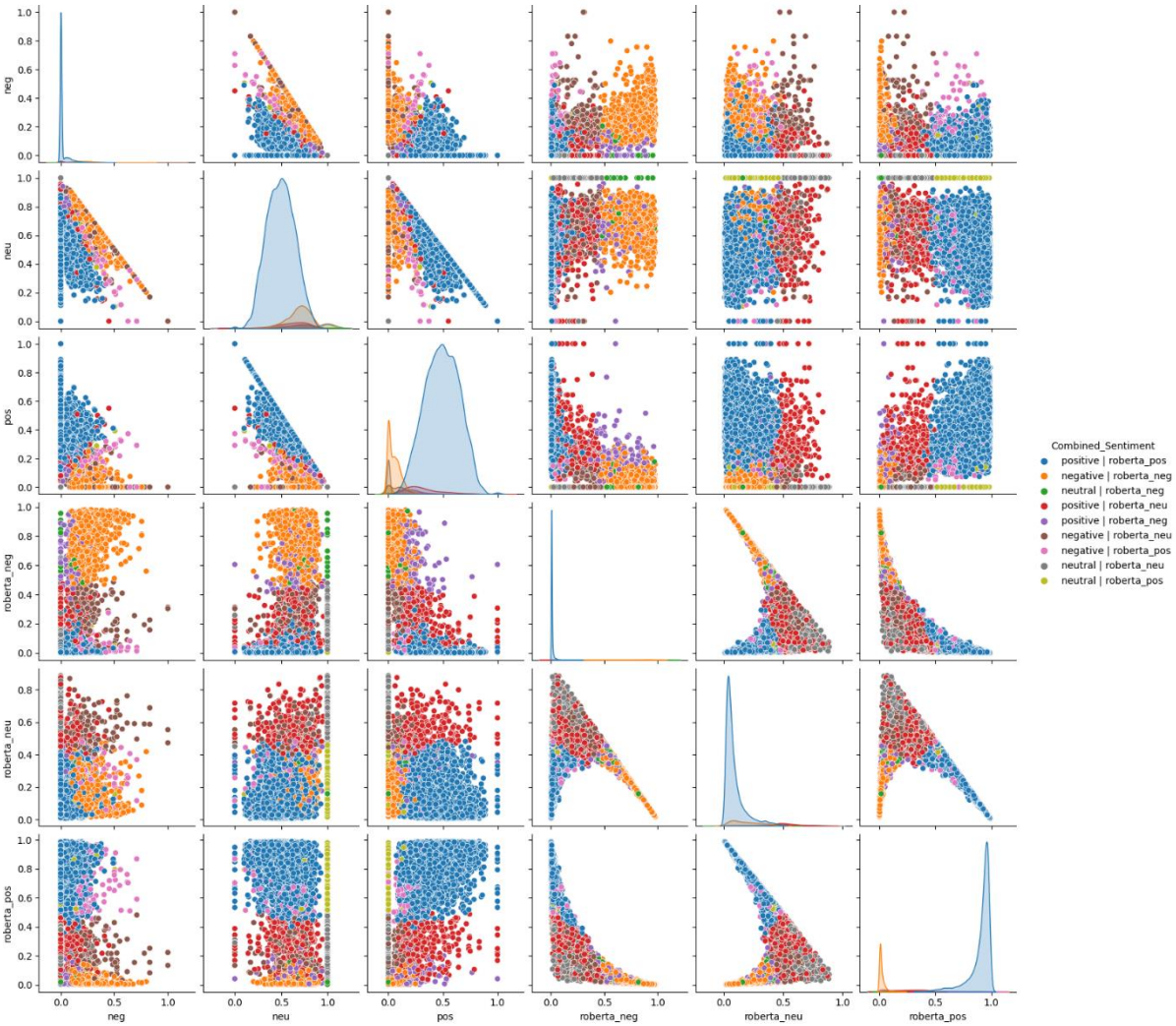

Trends in Positive Reviews Across Columns
value for money positive
garage service positive
ease of booking positive
location positive
length of fitting positive
delivery punctuality positive
garage service negative
tyre quality positive
mobile fitter positive
delivery punctuality negative
Number of Reviews
3500
3000
2500
2000
1500
1000
500
0
Column0
Column1
Column2
Column3
Column4
Column5
Column6
Column7
Column8
Columns

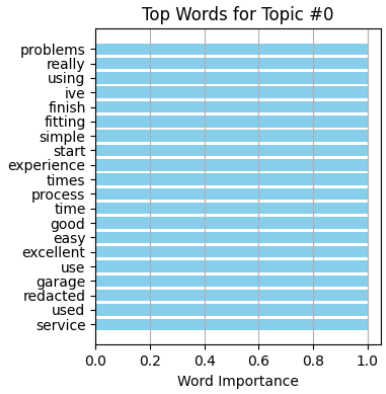

Sentiment Analysis
12000
10000
8000
6000
4000
2000
0
Count
Positive Sentiments
Negative Sentiments
Sentiment Polarity

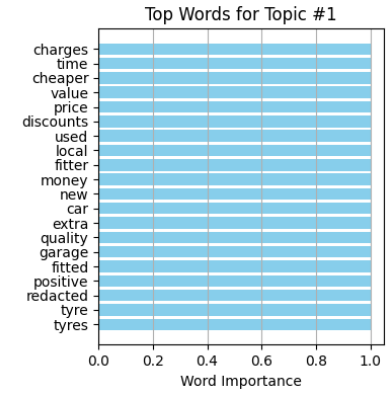

Sentiment Analysis Row Counts
7000
6000
5000
4000
3000
2000
1000
0
Count of Rows
At least one positive sentiment
At least one negative sentiment
Neither positive nor negative sentiment
Sentiment Presence

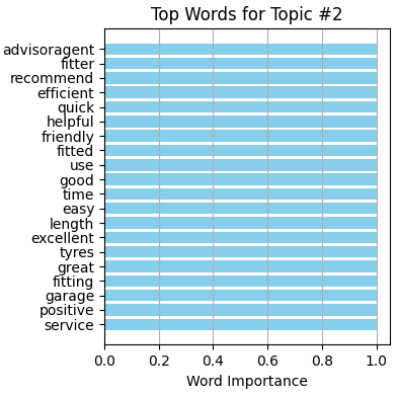

VADER Sentiment Distribution
Negative
11.8%
Neutral
2.7%
Positive
85.5%

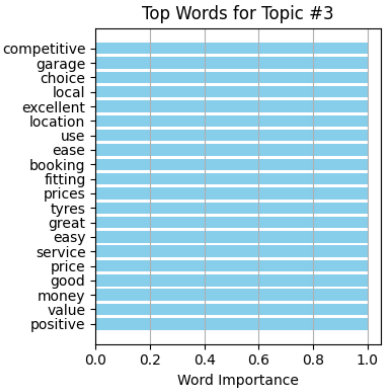

RoBERTa Sentiment Distribution
Negative
10.2%
Neutral
7.1%
Positive
82.7%

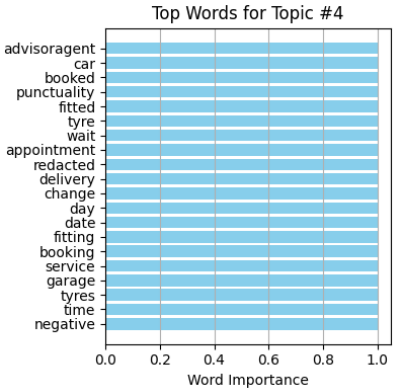

neg
neu
pos
roberta_neg
roberta_neu
roberta_pos
Combined_Sentiment
positive | roberta_pos
negative | roberta_neg
neutral | roberta_neg
positive | roberta_neu
positive | roberta_neg
negative | roberta_neu
negative | roberta_pos
neutral | roberta_neu
neutral | roberta_pos

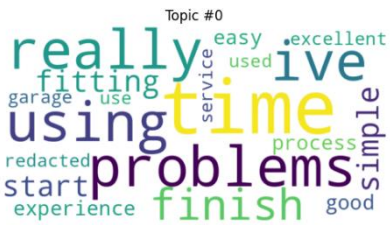

Top Words for Topic #0
problems
really
using
ive
finish
fitting
simple
start
experience
times
process
time
good
easy
excellent
use
garage
redacted
used
service
0.0
0.2
0.4
0.6
0.8
1.0
Word Importance

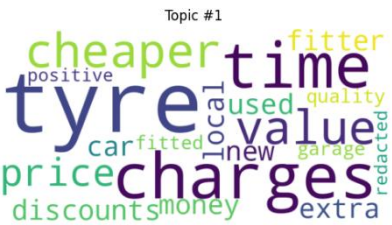

Top Words for Topic #1
charges
time
cheaper
value
price
discounts
used
local
fitter
money
new
car
extra
quality
garage
fitted
positive
redacted
tyre
tyres
0.0
0.2
0.4
0.6
0.8
1.0
Word Importance


Top Words for Topic #2
advisoragent
fitter
recommend
efficient
quick
helpful
friendly
fitted
use
good
time
easy
length
excellent
tyres
great
fitting
garage
positive
service
0.0
0.2
0.4
0.6
0.8
1.0
Word Importance


Top Words for Topic #3
competitive
garage
choice
local
excellent
location
use
ease
booking
fitting
prices
tyres
great
easy
service
price
good
money
value
positive
0.0
0.2
0.4
0.6
0.8
1.0
Word Importance

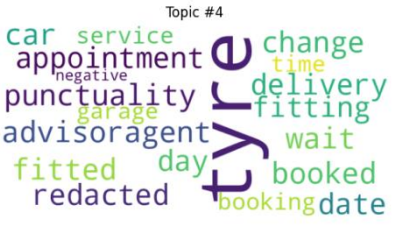

Top Words for Topic #4
advisoragent
car
booked
punctuality
fitted
tyre
wait
appointment
redacted
delivery
change
day
date
fitting
booking
service
garage
tyres
time
negative
0.0
0.2
0.4
0.6
0.8
1.0
Word Importance


Topic #0
really
easy
excellent
used
ive
fitting
service
garage
use
using
time
simple
process
redacted
problems
start
experience
finish
good


Topic #1
cheaper
fitter
time
positive
tyre
local
used
quality
value
car
fitted
new
garage
redacted
price
charges
discounts
money
extra


Topic #2
helpful
positive
efficient
garage
advisoragent
excellent
length
time
fitter
quick
fitting
friendly
fitted
great
use
easy
good
recommend
service
tyres


Topic #3
location
value
money
fitting
competitive
tyres
service
use
great
excellent
goodbooking
positive
local
price
ease
garage
choice
easy


Topic #4
car
service
change
appointment
negative
time
tyre
delivery
punctuality
fitting
garage
advisoragent
wait
fitted
day
booked
redacted
booking
date

Quality and completeness of code

1. Functionality: The algorithm counts dataset sentiments. It offers powerful sentiment analysis and visualizations using pretrained algorithms.

2. Clarity and Readability: The code is well-structured with step-by-step comments. This simplifies the analyzing procedure.

3. Advanced Techniques: Pretrained models like VADER and RoBERTa improve sentiment analysis by using advanced categorization methods.

Improvement Ideas

1. Subtheme Identification: The code should provide techniques to extract and classify subthemes to completely identify them and their attitudes. Possible methods include: - Text extraction for subtheme identification. Models can be used such as BERT, FLAN-T5

2. Model Fine-tuning: Pretrained models are reliable, but fine-tuning them on the dataset improves accuracy and relevance.

3. Evaluation Metrics: Precision, recall, and F1-score can quantify sentiment classification accuracy and performance.