جامعة أم القرى
UMM AL-QURA UNIVERSITY

Data Analysis 2

COURSE PRESENTER
(DR. Omima Fallatah)

SUBMITTED BY:

Name	ID
Jood Mohmmed Algarni	444005790
Noura Nawar Alhuthali	444001743

DEPARTMENT OF (DATA SCIENCE)
COLLEGE OF COMPUTERS
UMM AL-QURA UNIVERSITY

Data

The Yelp reviews polarity dataset categorizes reviews based on star ratings: stars 1 and 2 are labeled as negative (class 1), while stars 3 and 4 are labeled as positive (class 2). The dataset consists of 560,000 training samples and 38,000 testing samples, with 280,000 training samples and 19,000 testing samples for each polarity.
The `train.csv` and `test.csv` files contain the data in comma-separated values format, featuring two columns: the class index and the review text. Review texts are enclosed in double quotes, and internal double quotes are escaped using two double quotes. New lines are represented as "\n".

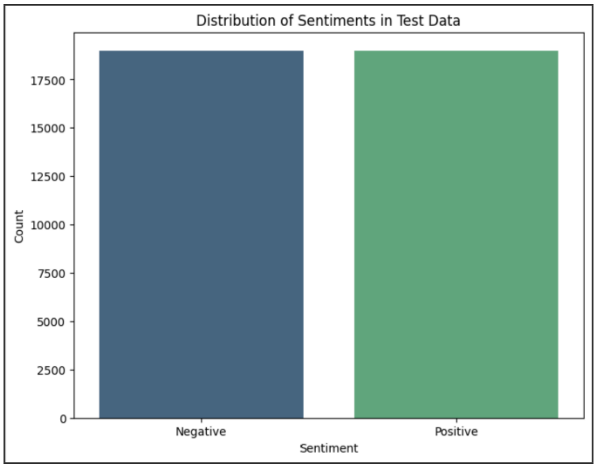
Defining Objectives

The primary Objective her is to classify Yelp reviews into negative (class 1) or positive (class 2) based on the sentiment expressed in the text this analysis can provide insights into customer opinions, helping businesses improve their products and services.

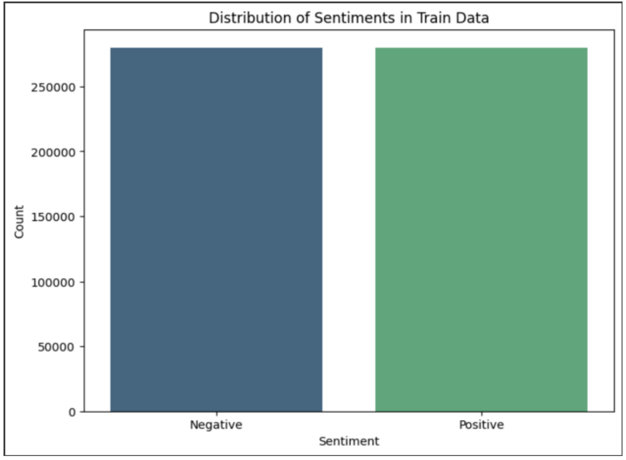
Data Exploration

The data was downloaded via Google Drive and displayed the first five lines of both files, then displayed the number of negative and positive classes and analysed the length of the revisions to better understand the data set, and this can be useful for understanding the overall distribution pattern and identifying any notable peaks, trends, or outliers in the data.

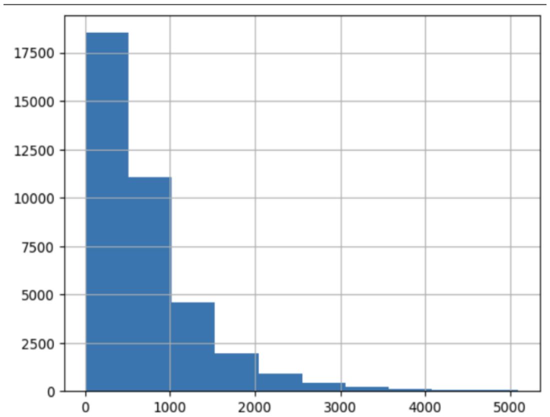
For test data:



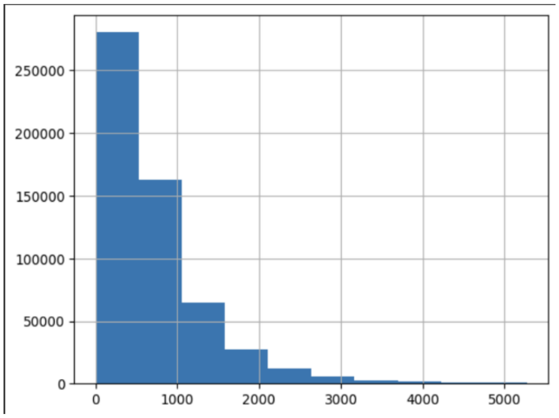
For training data:



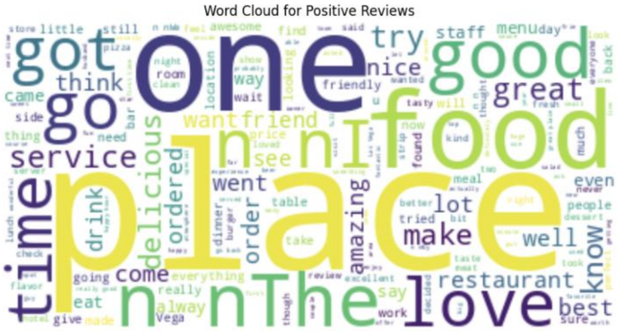
analysed the length for test:



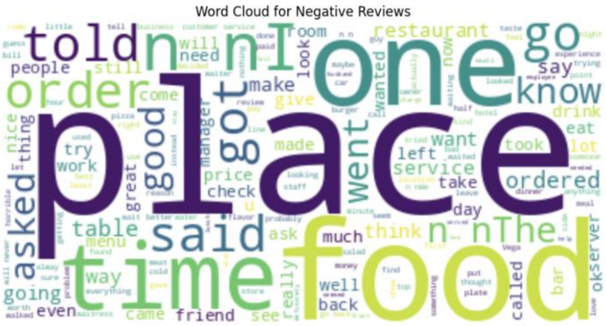
analysed the length for train:



The most frequent positive words:



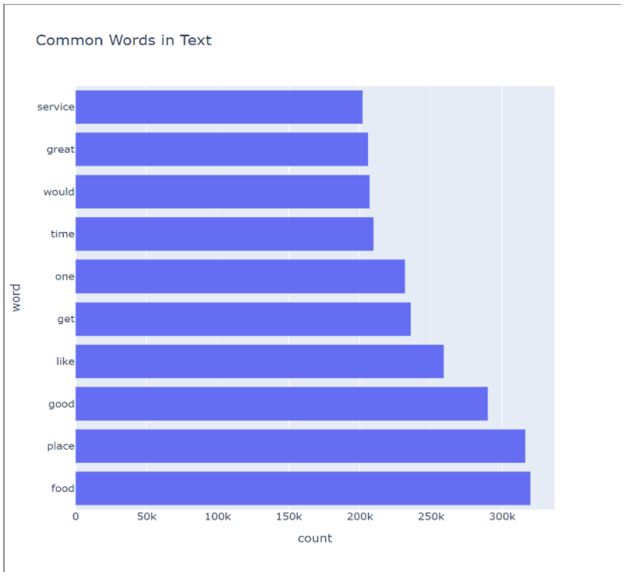
The most frequent negative words



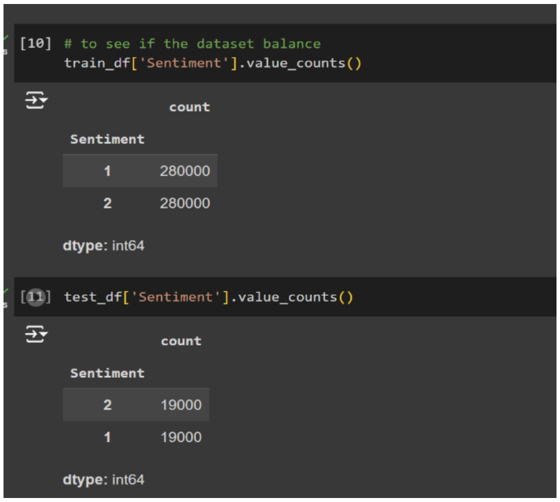
Data Processing

We changed the names of the columns to simple and expressive names that can be dealt with and then checking the data if it contains missing data or not and the result was that there is no missing data , and tokenization to Split the text into individual words or tokens, then remove stop words common words that may not contribute significant meaning (e.g., "and", "the", etc.), and stemming/lemmatization reduce words to their base or root form, and then we showed the 10 most repeated words to get us a background on what review is, making sure if the data is balanced or not and the result was that it was balanced.

the most frequency word:



Balance words in both training and test data:



```
[10] # to see if the dataset balance
     train_df['Sentiment'].value_counts()

              count
     Sentiment
         1    280000
         2    280000

     dtype: int64

[11] test_df['Sentiment'].value_counts()

              count
     Sentiment
         2    19000
         1    19000

     dtype: int64
```

Feature Extraction

- Vectorization: Convert the cleaned text into numerical format using methods like:
    - Count Vectorization: Counts the occurrence of words.
    - TF-IDF Vectorization: Considers the frequency of words and their importance.

We use TF-IDF to extract the words then apply machine learning on them

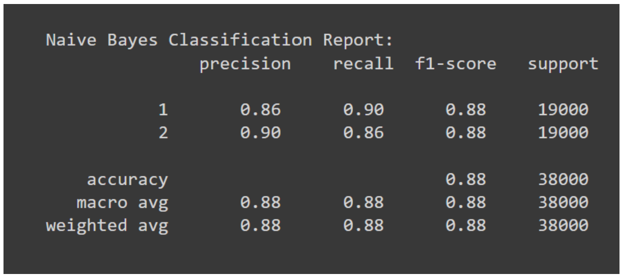
(<u>Naive Bayes</u>)

Choosing Models

Naive Bayes is a classification algorithm, which uses Bayes's theory of probability to predict an unknown category, we will use this algorithm to predict review if they are negative or positive.

Evaluating Model

To evaluate the model, we will check the (classification report) using actual and predicted values



```
Naive Bayes Classification Report:
              precision    recall  f1-score   support

           1       0.86      0.90      0.88     19000
           2       0.90      0.86      0.88     19000

    accuracy                           0.88     38000
   macro avg       0.88      0.88      0.88     38000
weighted avg       0.88      0.88      0.88     38000
```

1-Precision:

- Measures the ratio of correct predictions among all positive predictions made by the model.
- For class 1: 0.86 (86% of the predictions were correct).
- For class 2: 0.90 (90% of the predictions were correct).

2-Recall:

- Measures the ratio of correct predictions among all actual cases present in the data.
- For class 1: 0.90 (90% of the model's ability to classify emotions).
- For class 2: 0.86 (86% of the model's ability to classify emotions)

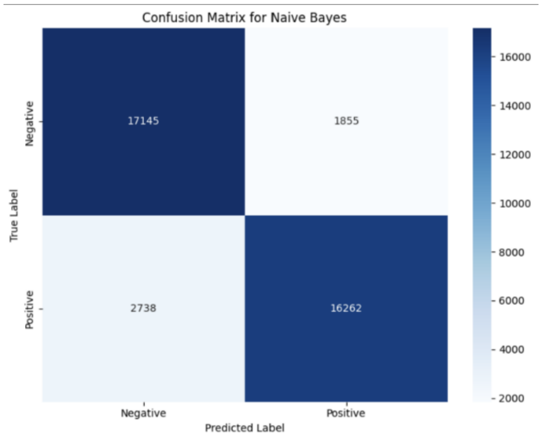
3-F1-Score:

- A metric that combines precision and recall, particularly useful when dealing with unbalanced classes.
- For class 1: 0.88 (a good value indicating a balance between precision and recall).
- For class 2: 0.88 (a good value indicating a balance between precision and recall).

Confusion Matrix

A confusion matrix is a tool used to evaluate the performance of a machine learning model in data classification. The matrix shows how the model classifies negative and positive emotions, helping to understand errors and correct predictions.

-True Negatives (TN): 17145 cases were correctly classified as negative.
-False Positives (FP): 1855 cases were incorrectly identified as having positive, while they negative.
-False Negatives (FN): 2738 cases were incorrectly identified as negative, while they are positive.
-True Positives (TP): 16262 cases were correctly classified as having positive.



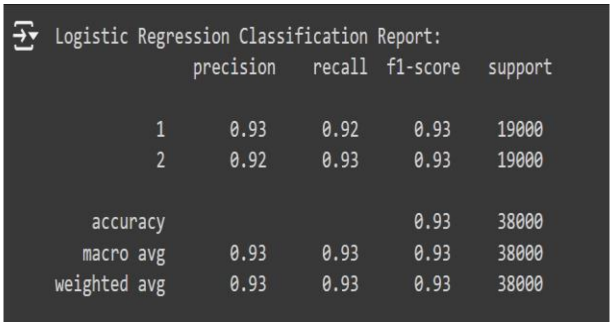
(Logistic Regression)

Choosing Models

Logistic regression is a statistical technique used to estimate the probability of a particular event, often applied in fields such as medicine and the social sciences. Using logistic regression, it is possible to infer the effect of sentiment analysis on a model's ability to analyse insights into customer feedback (negative or positive), helping to understand the factors contributing to the improvement of its products and services.

Evaluating Model

To evaluate the model, we will check the (classification report)using actual and predicted values



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

           1       0.93      0.92      0.93     19000
           2       0.92      0.93      0.93     19000

    accuracy                           0.93     38000
   macro avg       0.93      0.93      0.93     38000
weighted avg       0.93      0.93      0.93     38000
```

1-Precision:

- Measures the ratio of correct predictions among all positive predictions made by the model.
- For class 1: 0.93 (93% of the predictions were correct).
- For class 2: 0.92 (92% of the predictions were correct).

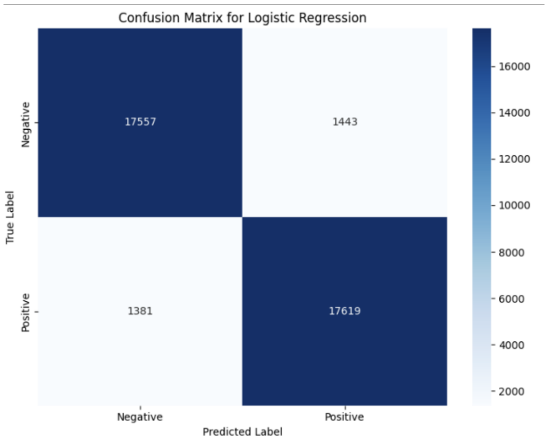
2-Recall:

- Measures the ratio of correct predictions among all actual cases present in the data.
- For class 1: 0.92 (92% of the model's ability to classify emotions).
- For class 2: 0.93 (93% of the model's ability to classify emotions)

3-F1-Score:

- A metric that combines precision and recall, particularly useful when dealing with unbalanced classes.
- For class 1: 0.93(a good value indicating a balance between precision and recall).
- For class 2: 0.93 (a good value indicating a balance between precision and recall).

Confusion Matrix



-True Negatives (TN): 17557 cases were correctly classified as negative.
-False Positives (FP): 1443 cases were incorrectly identified as having positive, while they negative.
-False Negatives (FN): 1381 cases were incorrectly identified as negative, while they are positive.
-True Positives (TP): 17619 cases were correctly classified as having positive.

Conclusion:

Based on the accuracy results of the two models, the logistic regression algorithm had a higher accuracy than the Naive Bayes model.
Additionally, the results from the logistic regression in the confusion matrix showed that the model made fewer incorrect predictions compared to Naive Bayes.
Therefore, the logistic regression algorithm is better in this analysis than Naive Baye.