Yelp Review Sentiment Dataset

COURSE PRESENTER:
(DR.Omaima A. Fallatah)

SUBMITTED BY:

Name	ID
Layan Adel Babkour	444002368
Reham Faisal Alsubhi	444003014

DEPARTMENT OF(INFORMATION SCIENCE-DATA SCIENCE)
COLLEGE OF COMPUTER AND INFORMATION SYSTEMS
UMM AL-QURA UNIVERSITY

Table of content

1.Introduction

Yelp is a leading platform dedicated to reviewing and rating local businesses, providing users with an effective way to discover and evaluate restaurants, cafes, hotels, and shops in their area. Since its founding in 2004, Yelp has become a go-to destination for consumers seeking reliable recommendations before making decisions. The site features an easy-to-use interface that allows users to search for places by location, type, or even business name, helping them find options that meet their needs. Yelp offers comprehensive information about each business, including star ratings, detailed user comments, and photos that reflect customer experiences. These features enhance individuals' ability to make informed decisions, and the ratings and reviews can significantly impact a business's reputation, making Yelp an essential tool for business owners to attract customers and improve their services. Additionally, Yelp data is utilized in market research and sentiment analysis to understand consumer trends and needs, further underscoring the platform's importance in the world of local commerce.

1.1.The importance of Understanding Customer Reviews

Knowing how to effectively utilize reviews, particularly on platforms like Yelp, is essential for both consumers and business owners. For consumers, reviews provide valuable insights into the quality and reliability of businesses, enabling informed decision-making.
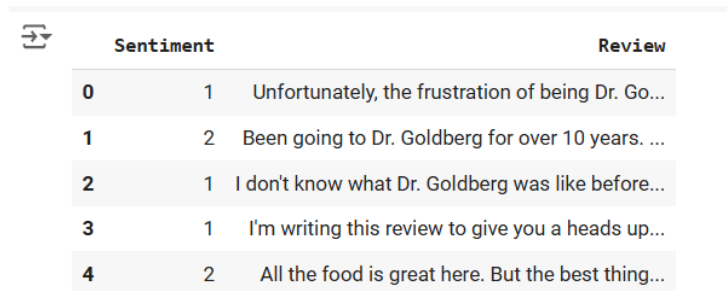
1.2.The our goals

- Determine the prevailing sentiments in user reviews (such as positive, negative) to understand customer satisfaction levels regarding different businesses.
- Apply preprocessing techniques such as tokenization and removing stop words.
- Learn and apply feature extraction techniques to data to improve the performance of machine learning models.

1.3.Link dataset

Yelp Review Sentiment Dataset (kaggle.com)

2.Exploratory Data Analysis

Datasets

	Sentiment	Review
0	1	Unfortunately, the frustration of being Dr. Go...
1	2	Been going to Dr. Goldberg for over 10 years. ...
2	1	I don't know what Dr. Goldberg was like before...
3	1	I'm writing this review to give you a heads up...
4	2	All the food is great here. But the best thing...

1. *Figure*: *Displays the Dataset.*

Libraries used

```python
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import accuracy_score, confusion_matrix
from sklearn.feature_extraction.text import TfidfVectorizer
```

2. *Figure*: *Shows the dataset libraries.*

3. *Figure*: *Displays attribute data type*

train.dtypes is used to determine the data type of each column in the dataset train.

4. *Figure*: *train.shape is used to obtain the dimensions (shape) of the data stored in the variable train*

train.shape is used to determine the dimensions of the dataset, specifically the number of rows and columns.
Number of rows: Represents the number of samples in the data (560,000).
Number of columns: Represents the number of features in the data (2).

```
train.isnull().sum()
```

                      0
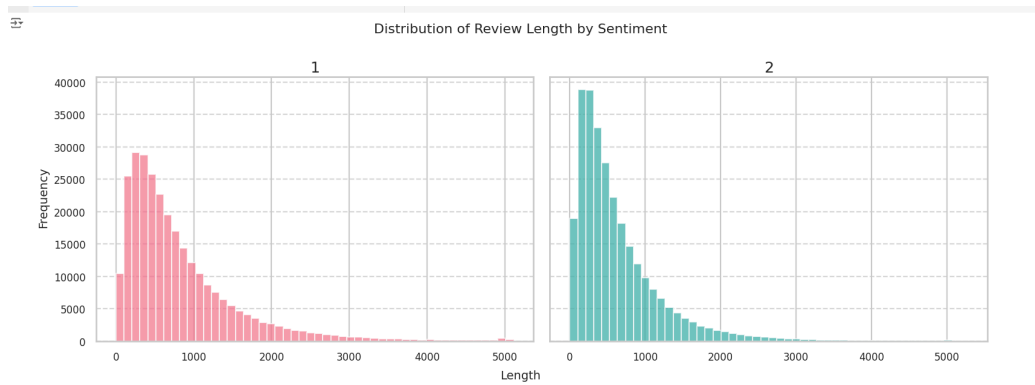   Sentiment   0
    Review   0

dtype: int64

5. *Figure : Verifies there are no missing values*

This function checks each element in the dataset and returns True if the value is missing (null) and False if it is not.

```
Longest Review:
 Francais plus bas.\n\n

Review Length: 5273


Shortest Review:
 A

Review Length: 1
```

6. *Figure : Display the longest and shortest length of the reviews*

We calculated the length of each review in the "Review" and added a new column called "Review_Length." The longest review was 5,273 characters long, while the shortest review was only 1 character.

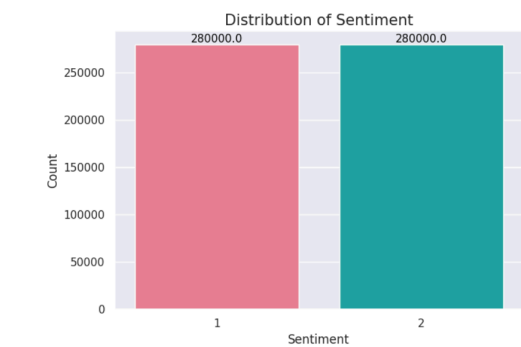6. *Figure* : *Display the normal distribution of reviews*

This chart illustrates the normal distribution of review lengths.


```
print(f"Sentiment value count: \n {train['Sentiment'].value_counts()}")
```

```
Sentiment value count:
 Sentiment
1    280000
2    280000
Name: count, dtype: int64
```

7. *Figure* : *Display the data is balanced or not*

This code shows us that our dataset is balanced.

3.Preprocessing:

In this code, we performed the preprocessing steps, which are:

Convert text to lowercase: We convert all letters in the text to lowercase to ensure consistency.
Remove URLs: We remove any URLs from the text.
Remove mentions and hashtags: We delete any mentions (@username) and hashtags (#topic) from the text.
Remove HTML tags: We eliminate any HTML tags that might be present in the text.
Expand contractions: We replace common contractions (e.g., "isn't" → "is not") to make the text clearer.
Remove punctuation: We remove all punctuation marks such as periods and commas.
Remove extra whitespaces: We eliminate any redundant spaces between words.
Tokenize and lemmatize: We split the text into individual words (tokens) and apply lemmatization to reduce words to their base forms.

	Sentiment	Review			Review
0	1	Unfortunately, the frustration of being Dr. Go...	0		unfortunately frustration dr goldberg patient ...
1	2	Been going to Dr. Goldberg for over 10 years. ...	1		going dr goldberg 10 year think one 1st patien...
2	1	I don't know what Dr. Goldberg was like before...	2		know dr goldberg like moving arizona let tell ...
3	1	I'm writing this review to give you a heads up...	3		im writing review give head see doctor office ...
4	2	All the food is great here. But the best thing...	4		food great best thing wing wing simply fantast...
Before removing stop words			**After removing stop words**		

4.Feature extraction:

The TF-IDF (Term Frequency-Inverse Document Frequency) technique is a text analysis tool used to measure the importance of words in specific documents within a larger set of documents. Term Frequency (TF) calculates how often a word appears in a document by dividing the number of times the word appears by the total number of words in the document. In contrast, Inverse Document Frequency (IDF) assesses the uniqueness of a word across a collection of documents by taking the natural logarithm of the total number of documents divided by the number of documents containing that word. This approach helps to reduce the weight of common words and highlight the most informative terms. By combining these two components, TF-IDF assigns a weight to each word, facilitating the identification of important and distinctive words in texts.

We have used training and testing in the **TF-IDF** method.

5.Building model :

MultinomialNB Model :

```python
vectorizer = TfidfVectorizer()
X_train = vectorizer.fit_transform(train['Review'])
X_test = vectorizer.transform(test['Review'])
model = MultinomialNB()
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
```

9. *Figure:Build the model, we used MultinomialNB.*

We have created a multinomial Naive Bayes model for text classification using the training data in X_tfidf and the test data in X_testfidf. After evaluating the model, we found that the accuracy is 88%.

Accuracy: 0.8841842105263158
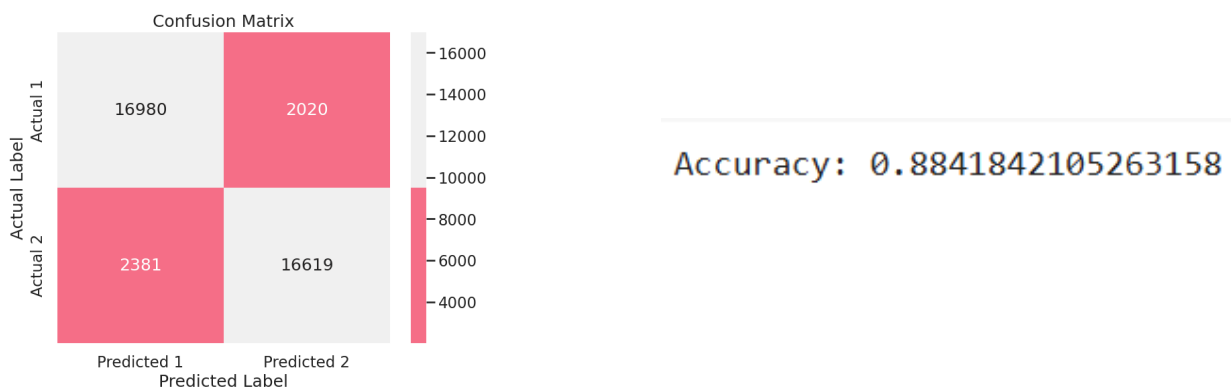
10. *Figure: Displays the confusion matrix and accuracy.*

6.Summary and Insight:

- Data balancing played a significant role in achieving this high accuracy.

- Techniques such as TF-IDF were used to extract key features, contributing to the identification of important and distinctive words in the texts.

- The model's good performance indicates the effectiveness of preprocessing steps (such as removing stop words and stemming) in improving the model's accuracy.

7.Conclusion:

The analysis of Yelp Review Sentiment provides valuable insights into the application of machine learning techniques for text data classification. Through meticulous preprocessing steps such as text normalization and the removal of irrelevant elements, combined with feature extraction techniques like TF-IDF, we were able to build an effective classification model using Multinomial Naive Bayes, achieving an accuracy of 88%. This model demonstrates a strong ability to predict user sentiment, highlighting the role of automated analysis in assessing customer satisfaction and enhancing business performance. The study emphasizes the importance of balanced datasets in constructing accurate and unbiased models, paving the way for broader applications of these techniques in consumer sentiment analysis.