

# Rapport de projet

## Sentiment Checker

Réalisé par :

- Najoua ZEFFATE
- Hajar AYADI

Sous la direction de :

Mr ALLAK Anas

Année universitaire 2022-2023

# Remerciement

C'est pour nous un plaisir autant qu'un devoir de remercier toutes les personnes qui ont pu prendre part, directement ou indirectement, à la réalisation du présent projet. Ainsi, nous tenons à exprimer nos plus sincères reconnaissances envers :

- Allah notre dieu qui nous a aidé à accomplir ce modeste travail.
- Mr. ALLAK Anas qui fait pour nous un encadrant attentif malgré ses charges nombreuses.

# Résumé

Ce projet est consacré à l'analyse des sentiments des clients à l'égard d'un site web de commerce en ligne en utilisant des techniques de NLP. L'objectif était de comprendre si les utilisateurs étaient généralement satisfaits ou insatisfaits du produit et de déterminer les principaux points de satisfaction ou de frustration.

Pour réaliser ce projet, nous avons collecté les commentaires sur le site web Booking à l'aide de l'outil de scrapage Web Scraper et les avons préparés en les nettoyant et en les tokenisant. Nous avons ensuite entraîné un modèle de classification automatique sur ces données pour prédire le sentiment des commentaires.

**Mots clés:** nlp, sentiment analysis, Scrapping, tokenization, web scraper, lemmatization, topic detection, language detection, text classification, deep learning, machine learning, selenium.

## Table des matières

I.	Scrapping : .....	7
II.	Cleaning data : .....	9
III.	Stocker les données dans une base de données relationnel : .....	12
IV.	Détection de la langue : .....	14
V.	Détection des sentiments : .....	17
VI.	Catégorisation des sujets : .....	18

# Introduction

Les avis et les commentaires des utilisateurs sur un site web peuvent être une source précieuse d'informations pour les entreprises et les organisations. Ils permettent de comprendre les points de satisfaction et de frustration des utilisateurs, ainsi que de déterminer les aspects à améliorer pour renforcer l'expérience utilisateur. Cependant, analyser manuellement un grand nombre de commentaires peut être fastidieux et coûteux. Heureusement, les techniques de traitement du langage naturel (NLP) peuvent aider à automatiser cette tâche.

Dans ce projet, nous avons utilisé des techniques de NLP pour analyser le sentiment des utilisateurs à l'égard du site web Booking. Nous avons collecté les commentaires des utilisateurs sur le site et avons utilisé un modèle de classification automatique pour évaluer le sentiment des commentaires. L'objectif de ce projet était de comprendre si les utilisateurs étaient généralement satisfaits ou insatisfaits du site et de déterminer les principaux points de satisfaction ou de frustration. Nous présenterons les résultats de notre analyse et les conclusions que nous avons tirées de ce projet dans les sections suivantes.

# Traitement du langage naturel (NLP)

Le traitement du langage naturel (NLP) est une branche de l'intelligence artificielle qui vise à permettre aux ordinateurs de comprendre et de traiter le langage humain de manière aussi naturelle que possible. Cela inclut la compréhension de la syntaxe, de la sémantique et de la structure du langage, ainsi que la capacité de traiter des textes et des conversations en langage naturel.

Le NLP peut être utilisé dans de nombreuses applications, notamment la traduction automatique, la reconnaissance de la parole, l'analyse de sentiments, la synthèse de la parole et la génération de texte. Il est également utilisé dans de nombreux domaines, tels que la recherche, la publicité, les relations avec les clients et les réseaux sociaux.

Le NLP repose sur des techniques de traitement du langage, de l'apprentissage automatique et de l'analyse de données pour extraire des informations et des connaissances utiles à partir de textes et de conversations. Il utilise également des algorithmes de traitement de la langue pour analyser et comprendre le langage naturel.

Le NLP est en constante évolution et de nouvelles techniques et outils sont développés régulièrement pour améliorer sa performance et sa précision. Cela permet d'appliquer le NLP à de nouvelles applications et de résoudre de nouveaux problèmes dans divers domaines.

## Sentiment analysis (NLP)

Sentiment analysis (or opinion mining) is a natural language processing (NLP) technique used to determine whether data is positive, negative or neutral. Sentiment analysis is often performed on textual data to help businesses monitor brand and product sentiment in customer feedback, and understand customer needs.

## I. Scrapping :

Le scrapping est le processus d'extraction de données à partir d'un site web.

Il peut être utilisé pour collecter des grandes quantités de données pour l'analyse de données, l'apprentissage automatique ou la création de bases de données...

Cela est généralement réalisé à l'aide d'un programme ou d'un script qui automatise le processus de demande de données au site web, l'analyse du contenu HTML ou XML résultant et l'extraction des données souhaitées.

Il existe plusieurs bibliothèques qui peuvent être utilisées pour le web scraping en Python, notamment : BeautifulSoup, Scrapy, Selenium.

Dans notre cas, on a utilisé selenium.

On va commencer par installer puis importer selenium et ses modules nécessaires :

```
!pip install selenium
```

```
from selenium import webdriver
from selenium.webdriver.chrome.options import Options
import time
chrome_options = Options()
```

Le module "webdriver" fournit l'interface pour contrôler le navigateur web, et le module "chrome.options" fournit des options pour personnaliser le comportement du navigateur Chrome. Le module "time" est utilisé pour ajouter des délais dans le script.

Créons maintenant une instance de contrôle du navigateur, naviguer vers l'URL par la méthode driver.get(), localiser et interagir avec les éléments de la page web avec la méthode driver.find\_element(), écrire dans une zone de texte avec send\_keys().

Dans ce script on va chercher les liens des Hotels qui se trouve à Marrakech dans le site Booking version anglais.

```

driver = webdriver.Chrome("chromedriver",options=chrome_options)
url = "https://www.booking.com/index.en-gb.html"
driver.get(url)
driver.get_screenshot_as_file("screenshot.png")
time.sleep(5)
XPath = "/html/body/div[1]/div[2]/div/form/div[1]/div[1]/div[1]/div[1]/label/input"

inputElement = driver.find_element("xpath",XPath)
time.sleep(5)

inputElement.send_keys('Marrakech')
time.sleep(5)

driver.find_element("xpath", '//button[@class="sb-searchbox__button"]').click()

liens = []
for i in range(4):
    links = driver.find_elements("xpath", "//a[@class='e13098a59f']")
    for l in links:
        liens.append(l.get_attribute("href"))

    driver.find_element("xpath", "//button[@aria-label='Next page']").click()
    time.sleep(5)

```

On va extraire pour chaque hôtel les informations suivantes : le nom, l'adresse, les propriétés ainsi les commentaires, les noms des auteurs et la note de chaque commentaire. Ces données vont être stocker dans la liste info.

```

driver2 = webdriver.Chrome("chromedriver",options=chrome_options)
info = []

for lien in liens:
    driver2.get(lien)
    time.sleep(5)
    try:
        nom = driver2.find_element("xpath", "//h2[@class='d2fee87262 pp-header__title']").text
        adress = driver2.find_element("xpath", "//span[@class='nhp_address_subtitle\njs-hp_address_subtitle\njq_tooltip\n']").text
        propriety = driver2.find_element("xpath", "//div[@class='hp_desc_important_facilities clearfix hp_desc_important_facilities']")
        driver2.find_element("xpath", '//span[contains(text(), "Read all reviews")]').click()
        time.sleep(5)
        nomcoms = []
        rates = []
        coms = []

        for i in range(20):
            nomcom = driver2.find_elements("xpath", "//li/div/div/div/div/div/div/span[@class='bui-avatar-block__title']")
            for l in nomcom:
                nomcoms.append(l.text)

            rate = driver2.find_elements("xpath", "//div[@class='bui-review-score__badge']")
            for l in rate:
                rates.append(l.text)

            com = driver2.find_elements("xpath", "//div[@class='bui-grid__column-9 c-review-block__right']")
            for l in com:
                coms.append(l.text)

            time.sleep(2)
            try:
                driver2.find_element("xpath", "//a[@aria-label='Next page']").click()
                time.sleep(5)
            except:
                break

        length = len(nomcoms)

        for i in range(length):
            pageinfo={}
            pageinfo['Hotel_Name']=nom
            pageinfo['Hotel_Address']=adress
            pageinfo['Hotel_Propriety']=propriety
            pageinfo['Auteur_Commentaire']=nomcoms[i]
            pageinfo['Commentaire']=coms[i]
            pageinfo['Rating']=rates[i]
            info.append(pageinfo)
        except:
            continue

```

A cette étape, on a fini la partie scrapping et nos données sont conservées dans info.



## II. Cleaning data :

Stocker les données dans une dataframe data :

```
Entrée [69]: import pandas as pd

data=pd.DataFrame(info,columns=['Hotel_Name','Hotel_Adress','Hotel_Propriety','Auteur_Commentaire','Commentaire','Rating'])
data
```

Out[69]:

	Hotel_Name	Hotel_Adress	Hotel_Propriety	Auteur_Commentaire	Commentaire	Rating
0	Riad Parfum d'Orient	7 Derb Gnaoua, Medina, 40000 Marrakech, Morocco	Most popular facilities\n1 swimming pool\nFree...	Dominique	Reviewed: 5 December 2022\nDans un Ryad zen et...	9.0
1	Riad Babouchta & Spa	38 derb rouda, Medina, 40000 Marrakech, Morocco	Most popular facilities\nFree WiFi\nSpa and we...	Jessica	Reviewers' choice Reviewed: 13 May 2022\nRelax...	10
2	Riad Babouchta & Spa	38 derb rouda, Medina, 40000 Marrakech, Morocco	Most popular facilities\nFree WiFi\nSpa and we...	Annabel	Reviewed: 21 December 2022\nFantastic stay - h...	9.0
3	Riad Babouchta & Spa	38 derb rouda, Medina, 40000 Marrakech, Morocco	Most popular facilities\nFree WiFi\nSpa and we...	Alexandra	Reviewed: 19 December 2022\nExceptional\n10\nL...	10
4	Riad Babouchta & Spa	38 derb rouda, Medina, 40000 Marrakech, Morocco	Most popular facilities\nFree WiFi\nSpa and we...	Al-muatasim	Reviewed: 19 December 2022\nExceptional\n10\nL...	10
...	...	...	...	...	...	...
15691	VILLA ILYANA	Résidence Garden City, Douar Jnane Tallaght, 4...	Most popular facilities\n1 swimming pool\nFree...	Chaimae	Reviewed: 22 July 2021\nVery good\n8.0\nThere...	8.0
15692	VILLA ILYANA	Résidence Garden City, Douar Jnane Tallaght, 4...	Most popular facilities\n1 swimming pool\nFree...	Wolfgang	Reviewed: 24 February 2020\nSuperb\n9.0\nThere...	9.0

Si on affiche un commentaire :

```
Entrée [78]: data["Commentaire"][15691]

Out[78]: 'Reviewed: 18 December 2022\nSuperb\n9.0\nLiked\n. The staff is friendly and p
rovide an exceptional service, they were always very helpful. The riad is beaut
ifully decorated and has delicious food.\nHelpful Not helpful'
```

On remarque qu'il faut enlever les éléments suivants :

```
'Reviewed: 18 December 2022\nSuperb\n9.0\nLiked\n. The staff is friendly and p
rovide an exceptional service, they were always very helpful. The riad is beaut
ifully decorated and has delicious food.\nHelpful Not helpful'
```

Pour cela, on a utilisé les expressions régulières pour supprimer :

- « Reviewed : » et « Reviewers' choice »

```
import re
comm = []
for com in data["Commentaire"]:
    string = com
    pattern = re.compile(r"((Reviewed:)|(Reviewers' choice))")
    result = pattern.sub(r"", string)
```

- La date de publication du commentaire

```
pattern = re.compile(r"(\d+)\s(\w+)\s(\d+)")
result = pattern.sub(r"", result)
```

- « (There are no comments available for this review) » et « (Show translation) »

```
pattern = re.compile(r"(There are no comments available for this review)")
result = pattern.sub(r"", result)
pattern = re.compile(r"(Show translation)")
result = pattern.sub(r"", result)
```

- “(1 person | 3 people) found this review helpful.”

```
pattern = re.compile(r"((\d+)\s(person|people)\s(found this review helpful.))")
result = pattern.sub(r"", result)
```

- “Property response:...”

```
pattern = re.compile(r"((Property response:)(\s(\w+))+)")
result = pattern.sub(r"", result)
```

- “(Helpful Not helpful)” et “(\n)”
- ...

```
pattern = re.compile(r"((Helpful Not helpful)|(\n))")
result = pattern.sub(r"", result)
pattern = re.compile(r"((((\d).(\d))|(\d+))(Liked))|(Disliked))")
result = pattern.sub(r"", result)
pattern = re.compile(r"((\d).(\d))")
result = pattern.sub(r"", result)
comm.append(result)
```

Pour les propriétés on utilise la fonction replace :

```
res = []
for sub in data["Hotel_Propriety"]:
    prop = sub.replace("Most popular facilities\n", "")
    prop = prop.replace("\n", " , ")
    res.append(prop)
```

Puis

enregistrer

les données dans un fichier csv :

```
data.to_csv('D:/master/1ere_annee/S3/partie1/BI/projet/Total_Booking_data_eng.csv', index=False, sep=',', quotechar='"')
```

	Hotel_Name	Hotel_Adress	Hotel_Propriety	Auteur_Commentaire	Commentaire	Rating
0	Riad Parfum d'Orient	7 Derb Gnaoua, Medina, 40000 Marrakech, Morocco	1 swimming pool , Free WiFi , Airport shuttle ...	Dominique	Dans un Ryad zen et beau à Marakkech.. - Le R...	9.0
1	Riad Babouchta & Spa	38 derb rouda, Medina, 40000 Marrakech, Morocco	Free WiFi , Spa and wellness centre , Airport ...	Jessica	Relaxing stay · friendly staff and a very re...	10
2	Riad Babouchta & Spa	38 derb rouda, Medina, 40000 Marrakech, Morocco	Free WiFi , Spa and wellness centre , Airport ...	Annabel	Fantastic stay - helpful staff, clean, comfor...	9.0
3	Riad Babouchta & Spa	38 derb rouda, Medina, 40000 Marrakech, Morocco	Free WiFi , Spa and wellness centre , Airport ...	Alexandra	Exceptional · Extremely kind staff, delicious...	10
4	Riad Babouchta & Spa	38 derb rouda, Medina, 40000 Marrakech, Morocco	Free WiFi , Spa and wellness centre , Airport ...	Al-muatasim	Exceptional · Everything · Nothing	10
...	...	...	...	...	...	...
15691	VILLA ILYANA	Résidence Garden City, Douar Jhane Tallaght, 4...	1 swimming pool , Free WiFi , Free parking , F...	Chaimae	Very good	8.0
15692	VILLA ILYANA	Résidence Garden City, Douar Jhane Tallaght, 4...	1 swimming pool , Free WiFi , Free parking , F...	Wolfgang	Superb	9.0
15693	VILLA ILYANA	Résidence Garden City, Douar Jhane Tallaght, 4...	1 swimming pool , Free WiFi , Free parking , F...	Anonymous	Superb	9.0
15694	VILLA ILYANA	Résidence Garden City, Douar Jhane Tallaght, 4...	1 swimming pool , Free WiFi , Free parking , F...	Anonymous	Very good	8.0
15695	VILLA ILYANA	Résidence Garden City, Douar Jhane Tallaght, 4...	1 swimming pool , Free WiFi , Free parking , F...	Anonymous	Superb	9.0

15696 rows × 6 columns

Ajouter une colonne qui va représenter la clé primaire des hôtels :

```
import pandas as pd
import numpy as np
values = df['Hotel_Name'].unique()
values = pd.Series(np.arange(len(values)), values)
df['Primary_Key'] = df['Hotel_Name'].apply(values.get)
```

df

	Hotel_Name	Hotel_Address	Hotel_Propriety	Auteur_Commentaire	Commentaire	Rating	Primary_Key
0	Riad Parfum d'Orient	7 Derb Gnaoua, Medina, 40000 Marrakech, Morocco	1 swimming pool , Free WiFi , Airport shuttle ...	Dominique	Dans un Ryad zen et beau à Marakkech... · Le R...	9.0	0
1	Riad Babouchta & Spa	38 derb rouda, Medina, 40000 Marrakech, Morocco	Free WiFi , Spa and wellness centre , Airport ...	Jessica	Relaxing stay · friendly staff and a very re...	10	1
2	Riad Babouchta & Spa	38 derb rouda, Medina, 40000 Marrakech, Morocco	Free WiFi , Spa and wellness centre , Airport ...	Annabel	Fantastic stay - helpful staff, clean, comfor...	9.0	1
3	Riad Babouchta & Spa	38 derb rouda, Medina, 40000 Marrakech, Morocco	Free WiFi , Spa and wellness centre , Airport ...	Alexandra	Exceptional · Extremely kind staff, delicious...	10	1
4	Riad Babouchta & Spa	38 derb rouda, Medina, 40000 Marrakech, Morocco	Free WiFi , Spa and wellness centre , Airport ...	Al-muatasim	Exceptional · Everything · Nothing	10	1
...	...	...	...	...	...	...	...
15691	VILLA ILYANA	Résidence Garden City, Douar Jnane Tallaght, 4...	1 swimming pool , Free WiFi , Free parking , F...	Chaimae	Very good	8.0	92
15692	VILLA ILYANA	Résidence Garden City, Douar Jnane Tallaght, 4...	1 swimming pool , Free WiFi , Free parking , F...	Wolfgang	Superb	9.0	92
15693	VILLA ILYANA	Résidence Garden City, Douar Jnane Tallaght, 4...	1 swimming pool , Free WiFi , Free parking , F...	Anonymous	Superb	9.0	92
15694	VILLA ILYANA	Résidence Garden City, Douar Jnane Tallaght, 4...	1 swimming pool , Free WiFi , Free parking , F...	Anonymous	Very good	8.0	92

Diviser les données entre deux dataframe :

➤ Pour les hôtels :

```
df_hotel = df[["Primary_Key", "Hotel_Name", 'Hotel_Address', 'Hotel_Propriety']].drop_duplicates()
```

df\_hotel

	Primary_Key	Hotel_Name	Hotel_Address	Hotel_Propriety
0	0	Riad Parfum d'Orient	7 Derb Gnaoua, Medina, 40000 Marrakech, Morocco	1 swimming pool , Free WiFi , Airport shuttle ...
1	1	Riad Babouchta & Spa	38 derb rouda, Medina, 40000 Marrakech, Morocco	Free WiFi , Spa and wellness centre , Airport ...
201	2	Jemaa ElFna Square Riad	Rue des Banques, Medina, 40000 Marrakech, Morocco	1 swimming pool , Free WiFi , Airport shuttle ...
220	3	Riad Lamy Marrakech	152 Derb Sidi Massaoud Bab Doukkala, Medina, 4...	Free WiFi , Airport shuttle , Parking , Family...
420	4	Riad Oriental De Marrakech	Derb Lgassaba Lakbir, Medina, 40030 Marrakech,...	1 swimming pool , Free WiFi , Airport shuttle ...
...	...	...	...	...
15206	88	Riad Sultan Suleiman	81 Derb El Hammam· Méchouar Kasbah, Kasbah, 40...	Free WiFi , Airport shuttle , Parking , Family...
15406	89	Riad Boussa	192 Derb Jdid Dabachi, Medina, Medina, 40400 M...	Free WiFi , Airport shuttle , Parking , Family...
15487	90	Riad Bensaid	1 Derb El Bir, Riad Zitoun Lakdim, Medina, 400...	Free WiFi , Airport shuttle , Parking , Family...
15628	91	Maison KA	N12 DERB SAADI AIN ITTI, 40000 Marrakech, Morocco	2 swimming pools , Free WiFi , Airport shuttle...
15653	92	VILLA ILYANA	Résidence Garden City, Douar Jnane Tallaght, 4...	1 swimming pool , Free WiFi , Free parking , F...

93 rows × 4 columns

➤ Pour les commentaires :

```
df_comment = df[["Primary_Key", "Auteur_Commentaire", 'Commentaire', 'Rating']]
```

```
df_comment
```

	Primary_Key	Auteur_Commentaire	Commentaire	Rating
0	0	Dominique	Dans un Ryad zen et beau à Marakkech.. · Le R...	9.0
1	1	Jessica	Relaxing stay · friendly staff and a very re...	10
2	1	Annabel	Fantastic stay - helpful staff, clean, comfor...	9.0
3	1	Alexandra	Exceptional · Extremely kind staff, delicious...	10
4	1	Al-muatasim	Exceptional · Everything · Nothing	10
...	...	...	...	...
15691	92	Chaimae	Very good	8.0
15692	92	Wolfgang	Superb	9.0
15693	92	Anonymous	Superb	9.0
15694	92	Anonymous	Very good	8.0
15695	92	Anonymous	Superb	9.0

15696 rows x 4 columns

Maintenant, enregistrons les données dans un fichier csv :

```
df_comment.to_csv('D:/master/1ere_annee/S3/partie1/BI/projet/Comment_Booking_data.csv', index=False)
df_hotel.to_csv('D:/master/1ere_annee/S3/partie1/BI/projet/Hotel_Booking_data.csv', index=False)
```

### III. Stocker les données dans une base de données relationnel :

1. Création de la connexion avec postgresql :

```
import psycopg2

conn = psycopg2.connect(database="Booking",
                        user='postgres', password='hajar',
                        host='localhost', port='5432'
)

conn.autocommit = True
cursor = conn.cursor()
```

2. Création de la table hotel :

```
sql = '''CREATE TABLE IF NOT EXISTS hotel(
        id int PRIMARY KEY,
        name text NOT NULL,
        adress text NOT NULL,
        propriety text NOT NULL
    );'''

cursor.execute(sql)
```

### 3. Insertion de données :

```
sql2 = '''copy hotel(
            id,
            name,
            adress,
            propriety)

FROM 'D:/master/1ere_annee/S3/partie1/BI/projet/Hotel_Booking_data.csv'
DELIMITER ','
CSV HEADER;'''

cursor.execute(sql2)
```

The screenshot shows the pgAdmin 4 interface. On the left, the 'hotel' table is selected under 'Tables (2)'. The main pane displays the query result for 'SELECT \* FROM public.hotel ORDER BY id ASC'. The result is a table with 17 rows and 4 columns: id, name, adress, and propriety.

id	name	adress	propriety
0	Riad Parfum d'Orient	7 Derb Gnaoua, Medina, 40000 Marrakech, Morocco	1 swimming pool , Free WiFi , Airport shuttle , Parki
1	Riad Babouchta & Spa	38 derb rouda, Medina, 40000 Marrakech, Morocco	Free WiFi , Spa and wellness centre , Airport shuttle
2	Jemaa ElFna Square Riad	Rue des Banques, Medina, 40000 Marrakech, Morocco	1 swimming pool , Free WiFi , Airport shuttle , Resta
3	Riad Lamyra Marrakech	152 Derb Sidi Massaoud Bab Doukkala, Medina, 40000 Marrakech, Morocco	Free WiFi , Airport shuttle , Parking , Family rooms ,
4	Riad Oriental De Marrakech	Derb Lgassaba Lakbir, Medina, 40030 Marrakech, Morocco	1 swimming pool , Free WiFi , Airport shuttle , Resta
5	Riad Assala	8, Derb Oualah Sidi Abdel Azziz Medina, Medina, 40000 Marrakech, Morocco	Free WiFi , Spa and wellness centre , Airport shuttle
6	Riad Jardin Des Sens & Spa	151 et 151 bis derb Jdid, Dabachi Medina, Medina, 40000 Marrakech, Morocco	3 swimming pools , Free WiFi , Spa and wellness ce
7	Riad Dar Jdidi	1bis Derb Doula Avenue Sidi Boulabada, Medina, 40000 Marrakech, Morocco	Free WiFi , Airport shuttle , Parking , Family rooms ,
8	Riad Alaka	35, Derb Alaka, Bab Doukala, Medina, 40000 Marrakech, Morocco	2 swimming pools , Free WiFi , Airport shuttle , Fam
9	RIAD SOKERA HOTEL RESTAU...	145 DERB JDID BAB DOUKALA, Medina, 40000 Marrakech, Morocco	1 swimming pool , Free WiFi , Spa and wellness cen
10	Riad Charai	S4, Diour Jdid, Zaouia Abbassia, Medina, 40000 Marrakech, Morocco	1 swimming pool , Free WiFi , Airport shuttle , Parki
11	Riad Jomana	62 Derb Aret Aouzal , Bab Doukkala, Medina, 40030 Marrakech, Morocco	Free WiFi , Airport shuttle , Parking , Family rooms ,
12	Riad Jnane Mogador	116, Riad Zitoun Kedim, Medina, 40000 Marrakech, Morocco	Free WiFi , Spa and wellness centre , Airport shuttle
13	Riad Al Wifak	20 Derb Lalla Bent El Amri, Medina, 40000 Marrakech, Morocco	1 swimming pool , Free WiFi , Family rooms , Restau
14	Riad Carina	76 Rue Berrima, Derb Touareg, 40000 Marrakech, Morocco	1 swimming pool , Free WiFi , Airport shuttle , Parki
15	Riad Saad	99 Sidi Bououqate, Medina, 44000 Marrakech, Morocco	Free WiFi , Airport shuttle , Parking , Non-smoking r
16	Dar Beja	37 Derb MANCHOURA , Mellah, 40000 Marrakech, Morocco	1 swimming pool , Free WiFi , Spa and wellness cen

### 4. Création et insertion dans la table comment :

```
create = '''CREATE TABLE IF NOT EXISTS comment(
            id SERIAL PRIMARY KEY,
            id_hotel INTEGER REFERENCES hotel(id),
            Auteur_Commentaire text NOT NULL,
            Commentaire text NOT NULL,
            Rating float NOT NULL
            );'''

cursor.execute(create)

insert = '''copy comment(
            id_hotel,
            Auteur_Commentaire,
            Commentaire,
            Rating)

FROM 'D:/master/1ere_annee/S3/partie1/BI/projet/Comment_Booking_data.csv'
DELIMITER ','
CSV HEADER;'''

cursor.execute(insert)
```

The screenshot shows the pgAdmin 4 interface. On the left, the 'comment' table is selected under 'Tables (2)'. The main pane displays a SQL query: `SELECT * FROM public.comment ORDER BY id ASC`. Below the query, the 'Data Output' tab shows the results of the query. The table has 5 columns: `id` (integer), `id_hotel` (integer), `auteur_commentaire` (text), and `commentaire` (text). The results show 18 rows of data, including hotel IDs, author names, and their respective comments.

id	id_hotel	auteur_commentaire	commentaire
1	1	Dominique	Dans un Ryad zen et beau à Marakech. - Le Ryad est très bien décoré avec beaucoup de goût. Le personnel Nadia et Rabie sont magnifiques/Serviables a
2	2	Jessica	Relaxing stay - friendly staff and a very relaxing atmosphere on the outskirts of the busy medina.
3	3	Annabel	Fantastic stay - helpful staff, clean, comfortable and convenient. - wonderful place. Great location, very clean and comfortable and helpful staff. Very prett
4	4	Alexandra	Exceptional - Extremely kind staff, delicious meal on site, beautiful and clean bathroom. Very comfortable!
5	5	Al-muatasim	Exceptional - Everything - Nothing
6	6	Aslie	Superb - The staff is friendly and provide an exceptional service, they were always very helpful. The riad is beautifully decorated and has delicious food.
7	7	Kristof	Exceptional - It was very comfortable, friendly staff and very close from centrum. Highly recommended !
8	8	Theodoros	Wonderful experience! - The staff was exceptional, very helpful and always kind and smiley. The location was great, not too far from the city centre, 15 mir
9	9	Baron	Nice stay. I would stay there again and/or I would recommend to others. - Breakfast was very nice. Staff was helpful and friendly. Property was very clean
10	10	Peter	Superb - Staff were amazing, very friendly, attentive and happy to have a laugh with you and nothing is any trouble! The building is also very attractive. The
11	11	Shaoting	Superb - The staffs went above and beyond to ensure our stay were nothing short of exceptional.
12	12	Julietta	Very nice hotel very typical. was very comfortable and with a great service and location - super clean and very friendly and helpful staff - all good :)
13	13	Ankit	Excellent stop over in Marrakech - Communication with Aurore (Property owner), Wifi access (inside room + common spaces), Room size and aesthetics, A
14	14	Atchara	It's was really good - The accommodation is very good, the staff take care during the stay very well. I recommend the dinner I like chicken lemon tagine It's
15	15	Sylvia	Exceptional - We spent 5 wonderful days in this really beautiful, well-designed and cosy Riad, just in the center of Medina. We had a walking distance to th
16	16	Rowin	Good stay with friendly staff - The staff was super friendly and everything was very clean. The riad is really beautiful, just like the rooms. - The beds are a t
17	17	Kevin	Best value Riad - Great value Riad, extremely close to central of the Medina. Highly recommend airport transfer the first time you arrive as to not get los
18	18	Susan	Good city location, comfortable and welcoming, good value for money - Property centrally located and entirely authentic. Very clean and comfortable, help

Total rows: 1000 of 15696 Query complete 00:00:02.179 Ln 1, Col 1

##### 5. Sélection de lignes :

```
select = '''select * from comment;'''
cursor.execute(select)
for i in cursor.fetchall():
    print(i)

conn.commit()
conn.close()
```

##### 6. Fermeture de la connexion :

```
# Fermez la connexion
conn.close()
```

## IV. Détection de la langue :

La détection de la langue en utilisant le deep learning consiste à utiliser un modèle de deep learning entraîné pour prédire la langue d'un texte en analysant les caractéristiques du texte. Cela peut être fait en utilisant une approche de classification supervisée, où le modèle est entraîné sur un grand ensemble de données de textes étiquetés avec leur langue respective.

Pour entraîner le modèle, le texte est d'abord pré-traité pour le mettre dans un format approprié pour le modèle, puis converti en vecteurs de nombres en utilisant un encodage de texte tel que le one-hot encoding ou le word embedding. Le modèle de deep learning peut être un réseau de neurones convolutionnel ou un réseau de neurones à long court terme (LSTM), qui est entraîné sur les vecteurs de texte et les étiquettes de langue correspondantes.

Une fois le modèle entraîné, il peut être utilisé pour prédire la langue de nouveaux textes en analysant les vecteurs de texte générés à partir de ceux-ci et en prédisant la langue la plus probable.

La précision de la détection de la langue peut être améliorée en utilisant d'autres techniques de traitement du langage naturel, telles que la reconnaissance de la langue en utilisant des n-grammes, en combinaison avec le deep learning.

#### 1. Modèle pré-entraîné :

Pour utiliser ce modèle, on a installé le package « langid » qui utilise une approche basée sur l'apprentissage automatique pour détecter la langue de textes de longueur quelconque.

```
!pip install langid
```

Une fois installé, on peut importer le module « langid » et utiliser les fonctions qu'il fournit pour détecter la langue de textes.

```
import langid
```

Créons une fonction « detect\_language() » qui utilise la fonction « langid.classify » qui prend en entrée un texte et renvoie en sortie la langue détectée ainsi qu'un score de confiance associé.

```
import langid

def detect_language(i):
    iso = ["af", "ar", "bg", "bn", "ca", "cs", "cy", "da", "de", "el", "en", "es", "et", "fa", "fi", "fr", "gu", "he", "hi", "hr",
          "Afrikaans", "Arabic", "Bulgarian", "Bengali", "Valencian", "Czech", "Welsh", "Danish", "German", "Greek", "English", "Spanis"]
    var = langid.classify(i)
    i = iso.index(var[0])
    return langue[i]
```

A cette étape, notre modèle est prêt à être utiliser, il faut importer les données à partir de notre base de données, puis prédire.

```
import psycopg2

conn = psycopg2.connect(database="Booking",
                        user='postgres', password='hajar',
                        host='localhost', port='5432'
)

conn.autocommit = True
cursor = conn.cursor()

select = '''select Commentaire from comment;'''
cursor.execute(select)
comm = []
for i in cursor.fetchall():
    comm.append(str(i))

conn.commit()
conn.close()
```



```

langue = []
for i in comm:
    langue.append(detect_language(i))

```

Le résultat de cette prédiction est le suivant :

```

import pandas as pd

df = pd.DataFrame({"Commentaire":comm,"Langue":langue})

df.sample(20)

```

	Commentaire	Langue
1324	(' perfect quick stay · All perfect · staff ',)	Latin
5857	(' séjour au top ! · la décoration, la sympath...	French
6241	(' Superb · colazione buona e ogni giorno dive...	Italian
5942	(' Superb · Tutto · La posizione Per raggiunge...	Italian
5601	(' מקום קסום חמאמם מגניב · מקסים, יחודי, נקי א	Hebrew
2064	(' Hôtel idéalement placé · L'accueil et la co...	French
15172	(' Fantastic · Heel erg gastvrij, toen we binn...	Dutch
13594	(' Superb · LocationHospitalityFoodCleanness · ...	English
14696	(' great short stay · position, great staff, c...	English
10368	(' Excellent place · The room was super cute a...	English
14608	(' Good · Very comfortable bed and the staff w...	English
8212	(' una estancia maravillosa · El Riad es preci...	Spanish
162	(' perfect stay · We liked everything. The bes...	English
15060	(' Exceptional · Everything about Riad Marchic...	English
7603	(' Lo mejor del viaje · Todo perfecto. Si vuel...	Spanish
7253	(' Exceptional · très beau Riad, avec un belle...	French
14349	(' Muy buen sitio para conocer Marrakech · EL ...	Spanish
10011	(' A perfect hotel for a stay in Marrakesh, cl...	English
208	(' Un plaisir! · L'emplacement central, la déc...	French



## 2. Modèle entraîné :

Dans cette partie nous avons entraîné notre modèle **MultinomialNB** de détection automatique du langage sur la base de données **basilb2s/language-detection** présentée dans le site web Kaggle via le lien suivant : <https://www.kaggle.com/datasets/basilb2s/language-detection?resource=download>.

Nous avons réparti notre data en une data pour entraînement et nous avons réservé 20% de la data originale pour le test. Pour éviter le surapprentissage, ainsi le modèle a atteint une accuracy de plus de 90% ce qui prouve sa performance.

Nous avons appelé notre modèle via le code présenté dans la figure ci-après :

```
[46] from sklearn.naive_bayes import MultinomialNB
      model = MultinomialNB()
      model.fit(x_train, y_train)

      MultinomialNB()
```

Ensuite nous avons défini la fonction **predict** pour la détection de la langue utilisée dans chaque commentaire de notre data sujette (data scrapée) comme l'illustre la figure suivante :

```
[50] def predict(text):
      x = cv.transform([text]).toarray() # converting text to bag of words model (Vector)
      lang = model.predict(x) # predicting the language
      lang = le.inverse_transform(lang) # finding the language corresponding to the predicted value
      return(lang[0])
      #print("The language is in",lang[0]) # printing the language
```

Au final nous avons ajouté à notre data la colonne **langtr** pour exploiter dans les manipulations ci-après.

```
[53] df4 = df3.assign(langtr=1nt)
```

## V. Détection des sentiments :

Pour cette partie de détection des sentiments nous avons utilisé un modèle pré-entraîné en appelant la librairie prédéfinie en langage python **vaderSentiment** de cette bibliothèque nous avons appelé **SentimentIntensityAnalyzer** pour appliquer le sentiment analysis sur notre data et plus précisément sur le champ "Commentaire", comme l'illustre la figure suivante:

```
[6] from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
```

Notre modèle donne des résultats comme suit :

```
[10] SA
      {'neg': 0.082, 'neu': 0.714, 'pos': 0.203, 'compound': 0.4939},
      {'neg': 0.017, 'neu': 0.949, 'pos': 0.033, 'compound': 0.3331},
      {'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0},
      {'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0},
```

Pour une bonne visualisation du score "Sentiment", nous avons ajouté quelques contraintes, comme suit :

Si le champ 'compound' est positif alors assigner la valeur 'compound'\*10 au champ **SentAC\_RAT** du commentaire.

Et si la valeur est négative là, nous avons distingué entre deux cas :

Si la valeur du champ 'compound' est supérieur à -0.5 dans ce notre algorithme va retourner la ('compound'\*10)+5 au champ *SentAC\_RAT* du commentaire.

Si non il va assigner 10-(('compound'\*10) au champ *SentAC\_RAT* du commentaire.

Comme l'illustre la figure suivante :

```
[77] SA1=[]
      sentiment = SentimentIntensityAnalyzer()
      for i in range(0,len(df)):
          text_1=df.loc[df.index[i], 'Commentaire']
          sent_1 = sentiment.polarity_scores(text_1)
          if sent_1['compound']>=0:
              SA1.append(sent_1['compound']*10)
          elif sent_1['compound']>=-0.5 and sent_1['compound']<0 :
              SA1.append((sent_1['compound']*10)+5)
          else:
              SA1.append((sent_1['compound']*10)+10)
```

Après affichage des trois premières lignes de notre dataframe nous avons obtenue le résultat figuré dans la figure suivante :

```
[108] df3.head(3)
```

	Hotel_Name	Hotel_Address	Hotel_Propriety	Auteur_Commentaire	Commentaire	Rating	SentAC	SentAC_Rat
0	Riad Parfum d'Orient	7 Derb Gnaoua, Medina, 40000 Marrakech, Morocco	1 swimming pool , Free WiFi , Airport shuttle ...	Dominique	Dans un Ryad zen et beau à Marakkech... Le R...	9.0	0.7345	7.345
1	Riad Babouchta & Spa	38 derb rouda, Medina, 40000 Marrakech, Morocco	Free WiFi , Spa and wellness centre , Airport ...	Jessica	Relaxing stay · friendly staff and a very re...	10.0	0.8718	8.718
2	Riad Babouchta & Spa	38 derb rouda, Medina, 40000 Marrakech, Morocco	Free WiFi , Spa and wellness centre , Airport ...	Annabel	Fantastic stay - helpful staff, clean, comfor...	9.0	0.9862	9.862

## VI. Catégorisation des sujets :

Dans cette catégorisation des sujets nous avons utilisé le Model *BERTopic* pour la langue anglaise.


Nous avons choisi de faire le processus de catégorisation des sujets que les commentaires en anglais, car plus de 70% de commentaire présent dans notre data, sont rédigé en langue anglaise.


Alors comme nous avons cité avant, nous avons utilisé BERTopic pour faire la catégorisation des sujets présent dans notre data.

```
[106] model = BERTopic(language="english")
```

Après avoir télécharger le modèle BERTopic comme l'illustre la figure ci-après :

```
▶ topics, probs = model.fit_transform(docs)
```


 Downloading: 100% 1.18k/1.18k [00:00<00:00, 47.3kB/s]

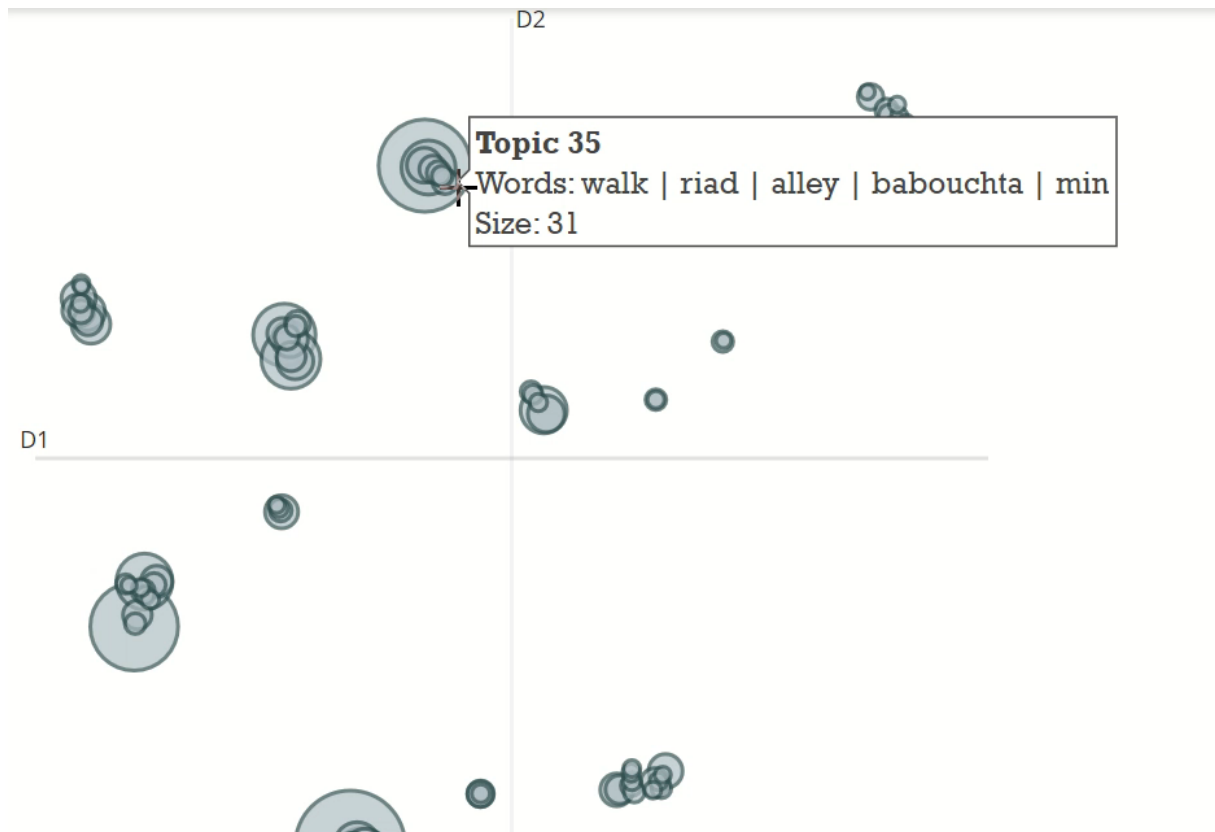

 Downloading: 100% 190/190 [00:00<00:00, 4.05kB/s]

Nous avons ensuite extrait les topics les plus fréquent comme suit :

```
[109] model.get_topic_freq()
```

	Topic	Count
0	-1	3027
1	0	586
2	1	413

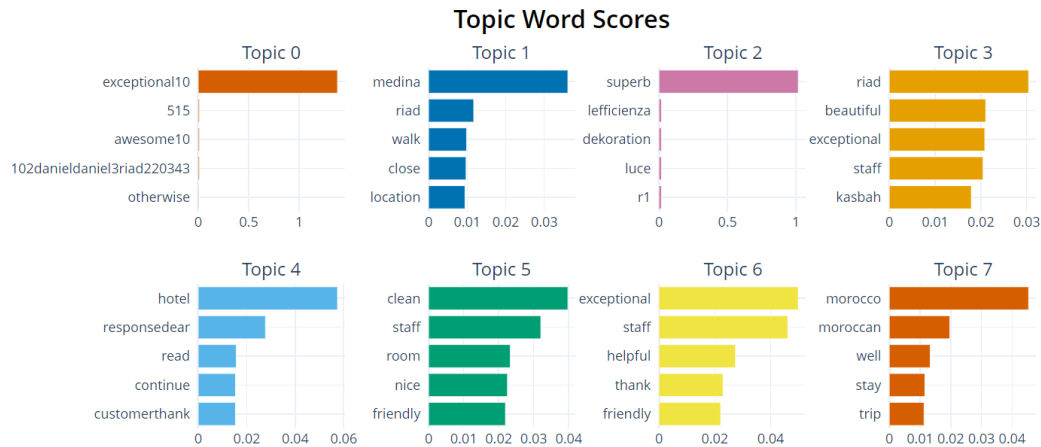
Pour une visualisation claire and clean nous avons visualiser nos topics via la méthode `visualize_topics()` pour voir les corrélation entre les topics. Comme suit :



Parmis les informations trivial que nous pu tirer de ce graphe c'est que :

La ville de Marrakech et l'hospitalité des gens sont fortement corrélées, et ce qui est vraiment le cas en réalité.

Nous avons aussi pu visualiser le score de chaque définissons un topic via la méthode `visualize_barchart()` comme l'illustre la figure suivante.



NB : le lien suivant vous amènera vers le code de la réalisation de ce projet :

<https://colab.research.google.com/drive/19gl8lapSHcm0Okub9V4Avl2BDe7rNVli?usp=sharing>

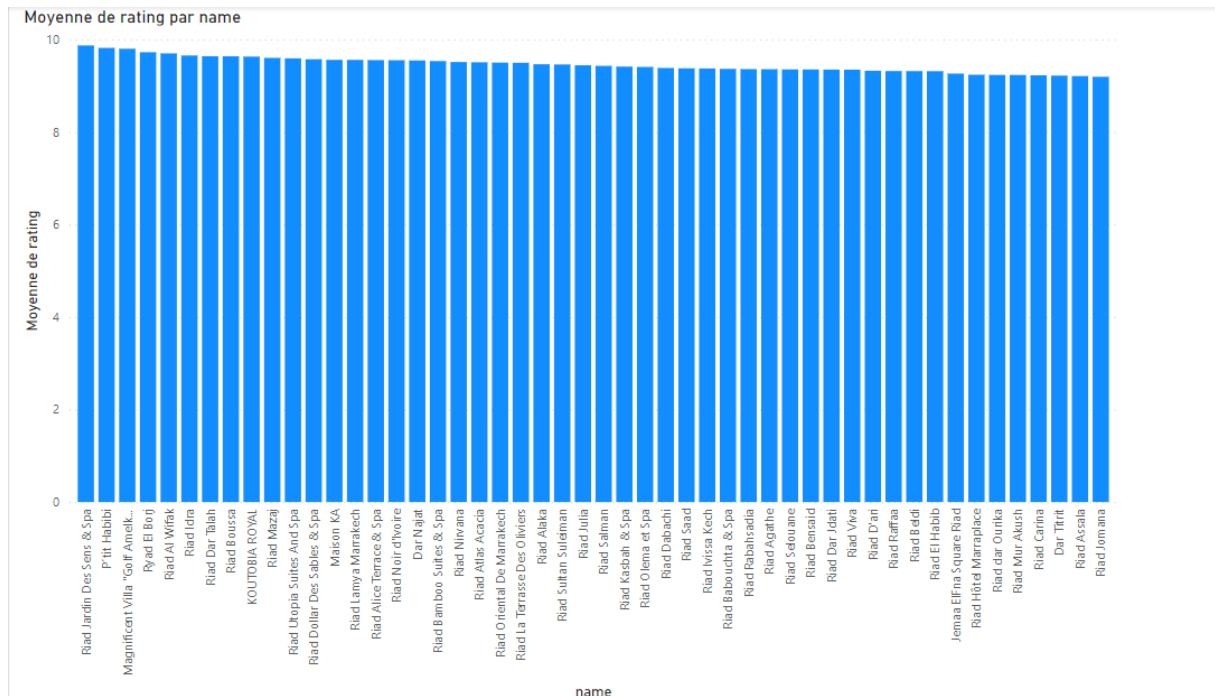
## VII. Partie visualisation par Power BI :

Premièrement nous avons stocker notre base donnée au niveau de PostgreSQL.

Et pour la partie visualisation nous avons utilisé l'outil power BI pour la réalisation d'un Dashboard.

Après avoir connecter power BI avec PostgreSQL, nous avons pu tirer le résultat illustré dans la figure suivante :

Afin de classier les meilleurs hôtels, on moyenne des ratings par hôtel



C'était une simple démonstration sur ce que nous pouvons faire avec power BI.

# Conclusion :

Durant ce projet, nous avons senti la puissance du web scrapping parce que nous avons pu tirer plusieurs informations de cette data, à savoir : la langue adoptée dans chaque commentaire ainsi le contexte de l'ensemble de commentaires. Et il reste plusieurs possibilités pour tirer plusieurs informations pertinentes de cette data, à titre d'exemple : prédire la satisfaction de futur client de différents pays.

En conclusion, l'analyse du sentiment des utilisateurs à l'aide de techniques de NLP peut être un moyen efficace de comprendre l'opinion des utilisateurs et de prendre des décisions en conséquence.

# Reference :

<https://stacklima.com/detecter-un-langage-inconnu-a-laide-de-python/>

<https://www.vennify.ai/bertopic-topic-modeling/>

<https://blog.deepgram.com/python-topic-modeling-with-a-bert-model/>

<https://towardsdatascience.com/end-to-end-topic-modeling-in-python-latent-dirichlet-allocation-lda-35ce4ed6b3e0>

<https://www.analyticsvidhya.com/blog/2021/03/language-detection-using-natural-language-processing/>

<https://www.kaggle.com/datasets/basilb2s/language-detection?resource=download>

<https://www.kaggle.com/code/dskswu/topic-modeling-bert-lda/notebook>

<https://www.analyticsvidhya.com/blog/2021/03/language-detection-using-natural-language-processing/>

<https://github.com/bhattbhavesh91/BERT-Topic-Modeling/blob/main/BERT-Topic-Modelling.ipynb>

<https://medium.com/analytics-vidhya/topic-modelling-using-lda-aa11ec9bec13>

<https://github.com/MaartenGr/BERTopic>

[https://github.com/susanli2016/NLP-with-Python/blob/master/LDA\\_news\\_headlines.ipynb](https://github.com/susanli2016/NLP-with-Python/blob/master/LDA_news_headlines.ipynb)

<https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24>

<https://www.geeksforgeeks.org/explanation-of-bert-model-nlp/>

<https://www.geeksforgeeks.org/fine-tuning-bert-model-for-sentiment-analysis/>

<https://www.analyticsvidhya.com/blog/2022/07/sentiment-analysis-using-python/>