UNIVERSITÉ MOHAMMED V

Faculté des Sciences de Rabat

Master Ingénierie de Données et Développement Logiciel

Natural Language Processing

Réalisé par : - Fadwa Saoiabi

- Fatima Zahra Atourabi

Encadré par : Mr Abdelhak Mahmoudi

Année universitaire : 2020-2021

I- Introduction :

Durant ce semestre, nous avons été initiés au Machine Learning et au NLP ainsi que leurs différentes techniques. Nous avons assimilé plusieurs algorithmes et méthodes utiles pour réaliser tout projet concernant le Natural Language Processing. Afin de matérialiser ces informations acquises et avoir une meilleure idée de ce qu'est le Machine Learning, notre professeur nous a donné libre recours de choisir un projet à réaliser concernant le NLP.

Pour notre part, notre choix s'est vite porté sur le **Sentiment Analysis**.

L'objectif de notre **projet** est de réaliser une analyse sentimentale des commentaires politiques présents dans notre jeu de données, du fameux site marocain Hespress. Dans un second temps, le but sera de parvenir à classifier les sentiments de ces commentaires selon qu'ils soient plutôt positifs, neutres ou négatifs à l'aide des différents modèles disponibles en Python.

II- Natural Language Processing :

Le traitement naturel du langage, ou Natural Language Processing (**NLP)** en anglais, est une technologie d'intelligence artificielle visant à permettre aux ordinateurs de comprendre le langage humain.

L'objectif de cette technologie est de permettre aux machines de lire, de déchiffrer, de comprendre et de donner sens au langage humain. D'importants progrès ont été effectués dans ce domaine au fil des dernières années, et le traitement naturel du langage est aujourd'hui exploité pour une large variété de cas d'usage…

- Quelles sont les différentes techniques de NLP ?

Les deux principales techniques utilisées pour le traitement naturel du langage sont **l'analyse syntaxique et l'analyse sémantique**.
L'analyse syntaxique consiste à identifier les règles grammaticales dans une phrase afin d'en déchiffrer le sens. L'analyse sémantique, quant à elle, consiste à **déchiffrer directement le sens d'un texte** en utilisant des algorithmes pour analyser les mots et la structure des phrases.

III - Sentiment Analysis :

L'analyse de sentiments (Sentiment Analysis ou Opinion Mining) est l'interprétation et la **classification des émotions** (positives, négatives et neutres) dans les données textuelles à l'aide des techniques d'analyse de texte. Cette analyse permet aux entreprises d'identifier l'opinion des clients à l'égard des produits, des marques ou des services dans les conversations et les commentaires en ligne.

L’analyse de sentiment se concentre sur **la polarité** (positive, négative, neutre), les sentiments et émotions (colère, joie, tristesse, etc.), et même sur les intentions (par exemple, intéressé contre non intéressé). Cela permet donc différents types d’utilisation de cette méthode d’analyse.

III- Implémentation et outils :

1- Langage de programmation :

Python est connu depuis longtemps comme un langage de programmation simple à maîtriser, du point de vue de la syntaxe. Il possède une communauté active et un vaste choix de bibliothèques et de ressources. On dispose donc d’une plate-forme de programmation capable d’être utilisée avec les technologies émergentes telles que l’apprentissage automatique et la Data Science.

2- Notebook :

La Data Science est itérative : il faut souvent tenter plusieurs approches et étudier les résultats avant de décider de la bonne façon de traiter un problème. C’est la raison pour laquelle les notebooks sont parfaitement adaptés à cette particularité. Un **notebook** est une interface web dans laquelle on peut taper du code Python, l'exécuter et voir directement les résultats, y compris une visualisation à l'aide de graphiques. Dans notre cas, nous avons choisi de travailler avec **Jupyter Notebook**.

3- Librairies utilisées :

Parmi les librairies Python que nous avons utilisées, on trouve :

• **NLTK :** C’est une librairie fondamentale pour la construction de programmes Python pour travailler avec des données de langage humain. Elle offre des interfaces faciles à utiliser sur des corpus ou ressources lexicales telles que WordNet, ainsi que des outils pour le traitement de texte, la classification, la tokenisation, le stemming, le balisage, l’analyse et le raisonnement sémantique.
• **SKLEARN :** C’est une librairie incontournable en Machine Learning et très bien documentée, destinée à **l'apprentissage automatique**. Elle comprend notamment des fonctions pour estimer des forêts aléatoires, des régressions logistiques, des algorithmes de classification ainsi que les machines à vecteurs de support.
• **Numpy :** C’est une librairie fondamentale pour effectuer des **calculs numériques** avec Python. Elle facilite grandement la **gestion des tableaux de données** et met à disposition également tout un **arsenal de fonctions pour effectuer des calculs mathématiques complexes** comme les fonctions trigonométriques ou encore les fonctions exponentielles et logarithmes.
• **Pandas :** C’est une **librairie très utilisée en data science** qui permet la manipulation et l'analyse des données. Elle propose en particulier des structures de données et des opérations de manipulation de tableaux numériques et de séries temporelles.

IV- Présentation du projet:

- ### Première analyse visuelle de notre Dataset :

Tout d’abord, voici un premier aperçu du jeu de données que nous avons à disposition. Comme nous pouvons l’apercevoir, cette data recensant au total 45857 de commentaires, contient deux colonnes intitulées “Comments” et “Score”.

Comments	Score
فليجتمعوا لما فيه خير ليبيا و ينهو هذا الصراع الذي عمر طويلا و دمر بلدا عزيزا علينا.	69
ابوى الرؤساء لن يحضروا يعني هناك ضغوط خارجية والجولة ستفشل بدونهما كما يتمناه أعداء الليبيين الذين يشلون حركة التفاوض اما عن قريب أو عن بعيد	-18
ا لامبراطورية المغربية العظمى لها تاريخ وشعب عريق واخواننا واشقاؤنا الليبيين يحبون ويثقون في المغرب , وعكس الجزائر اصبحت متورطة في استقبال القذافي وعائلته , والشعب ال	29
الحل الوحيد هو تقسيم ليبيا الى الشرق والغرب هذا هو الحل .أما المفاوضات فهي مجرد مضيعة الوقت انا مع تقسيم ليبيا أن شاء الله باذن الله	-87
نتمنى لاخوة الليبين ان يتوافقوا و يتفقوا على ما فيه خير ليبيا والليبين . ليبيا المدنية الديمقراطية الموحدة دولة المؤسسات الي يسيرها ابناءها بعيدا عن الاجندات الخارجية الي تريد اس	27
بعد الف جوله وجوله من المفاوضات ستبقى الحاله على ما عليها . ولكم في حماس وفتح لخير دليل على ذالك . في احد المحاولات العديده تم اقتيادهم للتوقيع على اتفاق في رحاب	0
الكبرائت اهون لديهم ان ينقائل الليبيون الى الابد من ان ينسب للمغرب اي فضل في الاخاء بينهم	17
اللهم وفقهم واصلح ذات بينهم وهذا سينخقق إن شاء الله إن إبتعد عنهم الشياطين خاصة الشياطين العرب الذين لا زالوا يظنون أن ليبيا الي يتحكم فيها القذافي لازالت قاءمة .إنكم د	3
كان عليهم ان يجتمعوا هده المرة بزعير	-1
Algeria is the tumor of the region	5
ليبيا لليبين بالتوفيق انشاء الله	3
لقد كان من البديهي عدم حضور السادة عقيلة صالح وخالد المشري مع حفظ الالقاب و تلويح الاخير بالإستقالة لأن حضورهما يعني قبولهما بتصفيتهما الجسدية من الجهات الي تد	2
صراحة الوضع معقد في ليبيا الشقيقة بحيث تتداخل فيه المصالح الاجنبية ودول الجوار ومع ذلك يمكن ان نتفاءل بحيث اللقاء الاول يزنبقة كسب الثقة بين المتحاورين وجعلهم يتد	0
نتمنى لاشقاءنا الليبيين كل الخير و ان يجدوا حلا نهائيا لازمتهم في بلدهم الثاني المغرب و يعم السلام و الازدهار بالتضامن و الوحدة في المنطقة المغاربية و الويل و العار لكل من له نية	2
المغرب لم يخذل أبدا إخوانه العرب!! صبر على تهورهم و طيشهم فكان حليما بقوة ثقته في نفسه!!! العرب لم يفهموا بعد أن الرزانة أم الحلول!!!! لهذا المغرب مستعد دائما لاستض	0
اللوي الصهيوني جند بعض الجيران للإبقاء على زعزعة الاستقرار في المنطقة لخدمت أجندتها التوسعية وقد اتضح هذا من خلال بعض التصريحات التي أدلى بها بعض القادة	0
الاقوياء والرجال الصحاح ،،،سوف يتناقشون يوم الاثنين القادم تحت قيادة المانيا وبحضور المشاركين في مؤتمر برلين ..هذا هو المؤتمر الذي سوف يذكره التاريخ وضيع المغرب المت	-3
نطلب من الله سبحانه وتعالى ان يبعد عن الأشقاء الليبيين بلاء كابرانات فرنسا. أمين يا رب العالمين	2
وزير الشغل في حكومتك ومن حزبك ووزير اخر من حكومتك وحزبك لم يسجلا موظفيهم ب cnss ولا شي اتخذ ضدهم ومازال عندك الوجه تقول بحال هاد الكلام!!!القيني على من	167
الهضرة فالتلفازة حاجا اما الواقع حاجا !!!!اخرى ا سي العثماني بغينا التطبيق ماشي الشفوي بغينا حقنا من الثروة بغينا نتخلصوا مزيان ماشي نعطونا الشياطة ديال 2500 درهم الي نتوم	140
حينما تسمع كلمة إصلاح اعلم أنها بالنسبة لحقوق الشعب ومكاسبه ما هي إلا إفساد، من قبيل الزحف على حق التقاعد وعلى القدرة الشرائية وعلى الوظيفة العمومية وعلى حقوق ال	58
خرجتو على الدراري اللي مبغاوش افراو ماخليتوهومش التعلمو الصنعة ما حدهم صغار بحكم انهم قاصرين حتى كبروا صنعة لا قرابة التلاوة بقرقوبي الحشيش ابرد ليهم لكتاف أصبحو	27
ما مصير مباراة المنتدبين القضائيين و التقنيين المتخصصين فالكهرباء و الماء .... لقد سئمنا وطال انتظارنا ؟؟؟؟؟	15

- ## Importation des librairies nécessaires :

  Avant de commencer, nous devons impérativement importer les bibliothèques qui nous seront utiles, à savoir NLTK, Pandas, Numpy, SkLearn etc..

```
import numpy as np
import pandas as pd
import nltk
import string
import re
import seaborn as sns
import matplotlib.pyplot as plt

from nltk import word_tokenize
from nltk.corpus import stopwords
from nltk.tokenize import wordpunct_tokenize

from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.pipeline import make_pipeline

from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import confusion_matrix,accuracy_score, classification_report
from sklearn.model_selection import train_test_split, GridSearchCV
```

Ensuite, nous avons importé notre dataset comme on le voit si bien sur la capture d'écran ci-dessous :

Entrée [2]:
```
data = pd.read_csv('comments.csv', sep=',', encoding='utf-8')
data.head()
```

Out[2]:

	comment	score
0	فليجتمعوا لما فيه خير لليبيا و يذهو هذا الصراع ...	69
1	ايوى الرؤساء لن يحضروا يعني هناك ضغوط خارجية و...	-18
2	ا لامبراطورية المغربية العظمى لها تاريخ وشعب ع...	29
3	الحل الوحيد هو تقسيم ليبيا الي الشرق والغرب هذ...	-87
4	نتمنى لاخوة الليبيين ان يتوافقوا و يتفقوا على م...	27

Entrée [3]:
```
data.shape
```

Out[3]:
```
(45857, 2)
```

- ## Pre-Processing :

  La première étape à chaque fois que l'on fait du NLP est de faire le nettoyage de nos données, autrement dit faire le **cleaning et le traitement** de notre Dataset. On a commencé d'abord par enlever les stopwords, ponctuations, diacritics, les mentions etc.. afin de se débarrasser des éléments qui ne serviront pas à grand-chose lors de notre Sentiment Analysis.

Etape 2: **Text pre-processing**

```
#List of arabic and english punctuations:

punctuations = '''`÷×؛<>_()*&^%][ـ،/:"؟.'{}~¦+|!”…“–ـ''' + string.punctuation

# Arabic stop words with NLTK :

stop_words = stopwords.words()

arabic_diacritics = re.compile("""
                             ّ    | # Shadda
                             َ    | # Fatha
                             ً    | # Tanwin Fath
                             ُ    | # Damma
                             ٌ    | # Tanwin Damm
                             ِ    | # Kasra
                             ٍ    | # Tanwin Kasr
                             ْ    | # Sukun
                             ـ     # Tatwil/Kashida
                         """, re.VERBOSE)

def preprocess(text):

# Remove punctuations :

    translator = str.maketrans('', '', punctuations)
    text = text.translate(translator)

# Remove diacritics :

    text = re.sub(arabic_diacritics, '', text)
```

```
# Remove_mention :

    text = re.sub(r'@\S+', '', text)

# Allow_only_ar :

    text = re.sub(r'[^\u0600-\u06ff\u0750-\u077f\ufb50-\ufbc1\ufbd3-\ufd3f\ufd50-\ufd8f\ufd50-\ufd8f\ufe70-\ufefc\uFDF0-\uFDFD]+

# Remove longation :

    text = re.sub("[إأآا]", "ا", text)
    text = re.sub("ي" ,"ى", text)
    text = re.sub("ء" ,"ؤ", text)
    text = re.sub("ء" ,"ئ", text)
    text = re.sub("ه" ,"ة", text)
    text = re.sub("ك" ,"گ", text)

# Remove stopwords :

    text = ' '.join(word for word in text.split() if word not in stop_words)


    return text

data['comment'] = data['comment'].apply(preprocess)
print(data.head(5))
```

Ainsi, nos données seront représentées comme suit :

		comment	score
0	69	فليجتمعوا خير ليبيا و ينهو الصراع عمر طويلا و ...	
1	18-	ايوي الرءساء يحضروا يعني ضغوط خارجيه والجوله س...	
2	29	ا لامبراطوريه المغربيه العظمي تاريخ وشعب عريق ...	
3	87-	الحل الوحيد تقسيم ليبيا الي الشرق والغرب الحل ...	
4	27	نتمني لاخوه الليبين ان يتوافقوا و يتفقوا علي خ...	

- ## Drawing a WordCloud :

Nous avons par la suite tracé un nuage de mots dans lequel on peut apercevoir **les mots les plus utilisés dans les différents commentaires des utilisateurs**. Plus la taille de la police d'un mot est grande, plus le nombre d'occurrences de celui-ci est important.

Etape 3: **Drawing a WordCloud**

```
import os
import codecs
from wordcloud import WordCloud
import arabic_reshaper
from bidi.algorithm import get_display


text_data = str(data['comment'])


# Make text readable for a non-Arabic library like wordcloud
text = arabic_reshaper.reshape(text_data)
text = get_display(text)

# Generate a word cloud image
wordcloud = WordCloud(font_path='NotoNaskhArabic-Regular.ttf').generate(text)

# Export to an image
img = wordcloud.to_file("arabic_example.png")
plt.imshow(img, interpolation='bilinear')
plt.axis("off")
```

```
(-0.5, 399.5, 199.5, -0.5)
```

- ## Sentiment Analysis with different techniques :

Concernant cette étape, nous avons commencé par classifier nos commentaires en se basant sur le score de ces derniers, ce qui nous a permis de déterminer si un commentaire est positif ou négatif.

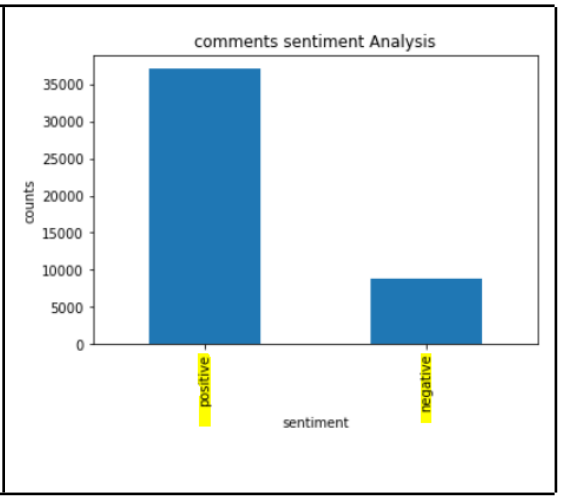
Sentiment Analysis with different techniques

```
#using score column
def getAnalysis(score):
    if score>=0:
        return 'positive'
    elif score<0:
        return 'negative'


data['Analysis'] = data['score'].apply(getAnalysis)
data
```

	comment	score	Analysis
0	فليجتمعوا خير ليبيا و ينهو الصراع عمر طويلا و ...	69	positive
1	ايوي الرءساء يحضروا يعني ضغوط خارجيه والجوله س...	-18	negative
2	... ا لامبراطوريه المغربيه العظمي تاريخ وشعب عريق	29	positive
3	... الحل الوحيد تقسيم ليبيا الي الشرق والغرب الحل	-87	negative
4	...تتمني لاخوه الليبين ان يتوافقوا و يتفقوا علي خ	27	positive
...	...	...	...
45852	...المغرب يخطو بالرياده الراعده سياسيا اجتماعيا ا	-4	negative
45853	...وزير الخارجيه الامريكي زار المغرب الاستقلال اق	3	positive
45854	...انا ارد علي سيد ابو وليد تعليقه خصوصا تبخيس مس	3	positive
45855	...ردا علي السيد فـءاد تكلم علي الضباط السامون للق	3	positive
45856	...ردا علي ابو صفاء علي علمي ان الرجل رغم انه دكت	1	positive

- Nous avons tracé par la suite un graphe qui montre clairement que notre Dataset contient + des avis positifs que négatifs.

```
data['Analysis'].value_counts()

plt.title('comments sentiment Analysis')
plt.xlabel('sentiment')
plt.ylabel('counts')
data['Analysis'].value_counts().plot(kind='bar')
plt.show()
```



• Multinomial Naïve Bayes :

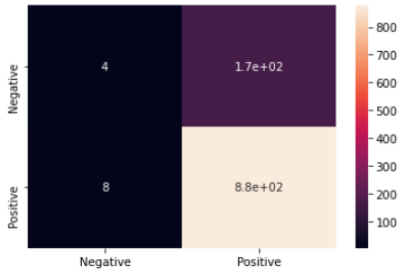
L'algorithme **Multinomial Naive Bayes** est une méthode d'apprentissage probabiliste principalement utilisée dans le traitement du langage naturel (NLP).
Cet algo est basé sur le théorème de Bayes, il prédit la balise d'un texte tel qu'un e-mail ou un article de journal et calcule la probabilité de chaque étiquette pour un échantillon donné, puis donne en sortie l'étiquette avec la probabilité la plus élevée.

```
X = CountVectorizer(analyzer = preprocess, dtype="uint8").fit_transform(data["comment"]).toarray()
y = data["score"].apply(lambda x: 'positive' if x >= 0 else 'negative')


X_train, X_test, y_train, y_test = train_test_split(X,y, test_size=0.023)

NB_classifier = MultinomialNB()
NB_classifier.fit(X_train,y_train)
```

```
MultinomialNB()
```

```
y_predict_test = NB_classifier.predict(X_test)
cm = confusion_matrix(y_test,y_predict_test)
sns.heatmap(cm, annot=True, xticklabels = ['Negative', 'Positive'],yticklabels=['Negative','Positive'])
print(classification_report(y_test,y_predict_test))
```

```
              precision    recall  f1-score   support

    negative       0.33      0.02      0.04       170
    positive       0.84      0.99      0.91       885

    accuracy                           0.84      1055
   macro avg       0.59      0.51      0.48      1055
weighted avg       0.76      0.84      0.77      1055
```



- ## Support Vector Machine :

  SVM est l'un des algorithmes d'apprentissage supervisé les plus populaires. Il est principalement utilisé pour les problèmes de classification ainsi que de régression. C'est même l'un des meilleurs modèles de classification.

Support Vector Machine (SVM)

```
#SVC :

clf = svm.SVC()
clf.fit(X_train, y_train)
pred=clf.predict(X_test)
print("SVC accuracy_score",accuracy_score(y_test, pred)*100)

cm = confusion_matrix(y_test,pred)
sns.heatmap(cm, annot=True, xticklabels = ['Negative', 'Positive'],yticklabels=['Negative','Positive'])

print(classification_report(y_test,pred))
```

```
SVC accuracy_score 83.88625592417061
```

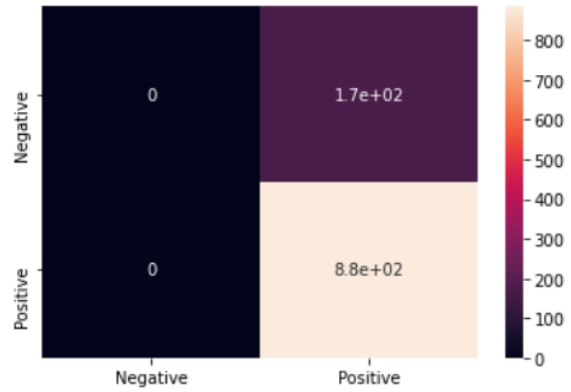
	precision	recall	f1-score	support
negative	0.00	0.00	0.00	170
positive	0.84	1.00	0.91	885
accuracy			0.84	1055
macro avg	0.42	0.50	0.46	1055
weighted avg	0.70	0.84	0.77	1055

- ## Logistic Regression :

  La régression logistique est une méthode statistique pour effectuer des classifications binaires. Elle prend en entrée des variables prédictives qualitatives et/ou ordinales et mesure la probabilité de la valeur de sortie en utilisant la fonction sigmoïd.

```
from sklearn.linear_model import LogisticRegression
logisticRegr = LogisticRegression()
logisticRegr.fit(X_train, y_train)
pred=logisticRegr.predict(X_test)
print("LogisticRegression accuracy_score",accuracy_score(y_test, pred)*100)

cm = confusion_matrix(y_test,pred)
sns.heatmap(cm, annot=True, xticklabels = ['Negative', 'Positive'],yticklabels=['Negative','Positive'])

print(classification_report(y_test,pred))
```

```
LogisticRegression accuracy_score 83.88625592417061
              precision    recall  f1-score   support

    negative       0.00      0.00      0.00       170
    positive       0.84      1.00      0.91       885

    accuracy                           0.84      1055
   macro avg       0.42      0.50      0.46      1055
weighted avg       0.70      0.84      0.77      1055
```



V- Conclusion :

Le Machine Learning, nouvel outil pour utiliser de la donnée, n'est pas encore déployé à son plein potentiel. Néanmoins, l'avancée technologique de cette méthode permet de découvrir de nouveaux cas d'études et de nouvelles opportunités.

Une fois comprise par le plus grand nombre d'entreprises, **l'analyse de sentiments** permettra de mieux comprendre les clients ainsi que de donner de nouvelles perspectives aux équipes afin d'obtenir un travail plus productif et de meilleure qualité.

Aujourd'hui, l'image qu'a une entreprise est très importante. Les répercussions d'une mauvaise image peuvent se manifester très rapidement, notamment avec les réseaux sociaux. C'est la raison pour laquelle il est recommandé d'utiliser l'analyse de sentiments, parce que ça peut aider à surveiller et à contrôler l'image d'un client ou d'une entité.