Université Ibnou Zohr-Agadir

Faculté Polydisciplinaire De Ouarzazate

Projet NLP pour MESURE Le Niveaux De Satisfaction Des Clients

Réalisé par :

GAJJA NOUR EDDIN

Encadé par :

CHARAF HAMIDI

Devant le jury :

Prof. **GOAU SALMA** : Professeur à la Faculté Polydisciplinaire de Ouarzazate

Prof. **CHARAF HAMIDI** : Professeur à la Faculté Polydisciplinaire de Ouarzazate

---

Année Universitaire : 2023-2024

Table des matières

Table des figures

DÉDICACE

A mes très chers Parents, Je dédie ce travail à mes parents, en reconnaissance de l'amour constant qu'ils m'ont témoigné, de leurs encouragements et du soutien inestimable qu'ils m'ont apporté tout au long de mes études. Aucun mot ni dédicace ne saurait suffire à exprimer pleinement le respect, la considération et l'affection que je ressens pour les sacrifices consentis par mes parents en vue de mon éducation et de mon bien-être. Que Dieu leur accorde santé, bonheur, prospérité et une longue vie, dans l'espoir que je puisse un jour leur rendre la joie qu'ils méritent tant

.

Remerciements

Nous adressons nos vifs et sincères remerciements à Monsieur Charafe Hamidi et Salma Goue, professeurs à la Faculté Polydisciplinaire de Ouarzazate, pour la qualité exceptionnelle de leur encadrement, leur soutien constant et leurs précieux conseils. Leur dévouement a grandement contribué à notre parcours académique. Nous exprimons également notre gratitude envers l'ensemble de nos professeurs et du personnel administratif de la faculté pour leur engagement et leur assistance précieuse. Nous souhaitons également exprimer notre reconnaissance à mes camarades de classe, dont le soutien et la collaboration ont enrichi notre expérience d'apprentissage et ont rendu cette aventure éducative encore plus mémorable

Introduction générale

Dans le paysage actuel des entreprises axées sur la satisfaction client, la capacité à comprendre les besoins, les préoccupations et les réactions des clients revêt une importance capitale. L'évolution rapide des technologies a ouvert la voie à de nouvelles méthodes d'analyse, notamment grâce au Traitement Automatique du Langage Naturel (TALN). Ce rapport se concentre sur l'application des techniques de TALN pour évaluer la satisfaction client à travers l'analyse des commentaires, des opinions et des sentiments exprimés par les clients dans diverses interactions, telles que les avis en ligne, les réseaux sociaux ou les enquêtes..

L'objectif principal de cette étude est d'explorer comment les avancées en matière de TALN peuvent être mises à profit pour comprendre plus efficacement les attentes et les niveaux de satisfaction des clients. Nous examinerons les méthodes utilisées pour collecter et analyser les données textuelles, en mettant l'accent sur les techniques de traitement du langage naturel telles que la reconnaissance des entités nommées, l'analyse de sentiment et la modélisation du langage.

Au travers de cette analyse approfondie, nous chercherons à démontrer comment les entreprises peuvent tirer parti des informations extraites des données textuelles pour améliorer leurs produits, services et interactions client. Enfin, nous discuterons des défis potentiels et des opportunités liées à l'application de ces techniques dans le domaine de la mesure de la satisfaction client, offrant ainsi des recommandations pour une mise en œuvre efficace et évolutive de ces pratiques dans les stratégies d'entreprise.

Et pour présenter ce rapport on suit un plan qui se compose de deux chapitres :

Chapitre 1 : Introduction générale

1.1 Contexte du projet

1.2 Objectif du projet

1.3 Problématique

Chapitre 2 : Méthodologies

2.1 Collecte de données

2.2 Description de la base de données

2.3 Nettoyage des données

2.4 Visualisation

2.5 Modélisation

2.6 Résultats du modèle

2.7 Interprétation des résultats

2.8 Langage utilisé

2.9 Les logiciels

Chapitre 3 : Conclusion

Chapitre 1

Introduction générale

1.1 Contexte du projet

Dans le cadre de notre projet axé sur le traitement du langage naturel (NLP) appliqué à la satisfaction client, l'objectif est d'exploiter les avancées technologiques pour analyser de manière exhaustive les retours et les commentaires des clients. En intégrant des techniques avancées de NLP, nous aspirons à dépasser les limites des approches conventionnelles, comme les enquêtes manuelles, en extrayant des insights pertinents et en temps réel à partir des données textuelles.

Ce projet vise à utilisé un modèle NLP sophistiqué capable de comprendre les nuances du langage humain, de détecter les sentiments exprimés dans les commentaires des clients, et d'appréhender les motifs sous-jacents. En adoptant cette approche, nous visons à obtenir une compréhension approfondie des besoins, des préoccupations et des points forts de nos clients, ce qui nous permettra d'améliorer continuellement la qualité de nos produits et services.

1.2 Objectif du projet

Notre projet a pour objectif principal d'intégrer l'analyse des sentiments, une application majeure du traitement du langage naturel (NLP), dans le domaine du marketing. Actuellement, l'analyse des sentiments est largement reconnue comme l'une des applications les plus populaires du NLP, notamment pour les spécialistes du marketing. Cette branche du NLP vise à décoder l'émotion et le ton d'un texte, permettant ainsi de le relier à une émotion, une opinion ou une attitude. L'analyse des sentiments joue un rôle essentiel en aidant les spécialistes du marketing à cartographier les émotions des clients à l'aide d'algorithmes complexes, ce qui leur permet d'offrir un soutien émotionnellement intelligent aux clients.

1.3 Problématique

Dans notre quotidien, les médias sociaux sont devenus une plateforme incontournable où de multiples commentaires émergent constamment. L'abondance de ces réactions rend l'analyse manuelle extrêmement chronophage. Face à cette réalité, le besoin de comprendre et de mesurer la satisfaction des clients à partir de ces commentaires devient impératif. C'est dans ce contexte que notre projet NLP prend tout son sens.

La problématique centrale de notre projet réside dans la nécessité de traiter efficacement et rapidement le flux massif de commentaires présents sur les médias sociaux. En utilisant le traitement du langage naturel (NLP), notre objectif est de mettre en place une méthodologie robuste permettant de mesurer la satisfaction des clients à partir de ces données. Cette approche permettra non seulement de gagner du temps, mais également d'obtenir des insights plus précis et en temps réel.

Ainsi, la question fondamentale que notre projet cherche à résoudre est la suivante : Comment utiliser le NLP pour analyser et mesurer la satisfaction des clients à partir des commentaires abondants sur les médias sociaux, tout en optimisant le temps consacré à cette tâche ? En répondant à cette problématique, nous visons à fournir une solution innovante qui permettra aux entreprises de mieux comprendre les sentiments de leurs clients et d'ajuster leurs stratégies en conséquence, tout en optimisant l'efficacité de l'analyse des commentaires sur les médias sociaux.

Chapitre 2

Méthodologies

2.1 Collecte de données

Importez le jeu de données :

Téléchargement des données et leur nettoyage.

```
In [2]: df = pd.read_csv('C:\\Users\\HP\\Desktop\\dataprojet\\Reviews.csv')
        df
```

Out[2]:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	Summary	Text
0	1	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	1	1	5	1303862400	Good Quality Dog Food	I have bought several of the Vitality canned d...
1	2	B00813GRG4	A1D87F6ZCVE5NK	dll pa	0	0	1	1346976000	Not as Advertised	Product arrived labeled as Jumbo Salted Peanut...
2	3	B000LQOCH0	ABXLMWJIXXAIN	Natalia Corres "Natalia Corres"	1	1	4	1219017600	"Delight" says it all	This is a confection that has been around a fe...

Voici les colonnes que j'ai dans la base de données. :

[1-3]

```
['Id', 'ProductId', 'UserId', 'ProfileName', 'HelpfulnessNumerator', 'HelpfulnessDenominator', 'Score', 'Time', 'Summary',
'Text']
```

FIGURE 2.1

2.2 Description du base de données

Ce jeu de données est composé d'avis sur des produits alimentaires de qualité provenant d'Amazon. Les données couvrent une période de plus de 10 ans, incluant tous les 500 000 avis jusqu'à octobre 2012. Les avis comprennent des informations sur le produit et l'utilisateur, des évaluations et un avis en texte brut. Il inclut également des avis de toutes les autres

catégories d'Amazon.

Index des lignes dans le DataFrame, de 0 à 568 453.
10 colonnes au total dans cet ensemble de données.

TABLE 2.1 – Informations générales

Id	Identifiant, `int64`.
ProductId	Identifiant du produit, `object`.
UserId	Identifiant de l'utilisateur, `object`.
ProfileName	Nom du profil, `object`.
HelpfulnessNumerator	Numérateur d'aide, `int64`.
HelpfulnessDenominator	Dénominateur d'aide, `int64`.
Score	Score, `int64`.
Time	Temps, `int64`.
Summary	Résumé, `object`.
Text	Texte, `object`.

TABLE 2.2 – Résumé des colonnes avec types de données

Informations sur les types de données (`dtypes`) : Nombre de valeurs non nulles et types de données pour chaque colonne.
Utilisation de la mémoire : 43.4+ MB.

TABLE 2.3 – Informations sur les types de données

2.3 Nettoyage des données :

Le nettoyage des données est essentiel pour le traitement automatique du langage, et pour cela, nous avons besoin d'effectuer le nettoyage des données.

Voici le nombre des colones vides.

Voici les nombre des lignes vides.

```
RangeIndex: 568454 entries, 0 to 568453
Data columns (total 10 columns):
 #   Column                  Non-Null Count   Dtype
---  ------                  --------------   -----
 0   Id                      568454 non-null  int64
 1   ProductId               568454 non-null  object
 2   UserId                  568454 non-null  object
 3   ProfileName             568428 non-null  object
 4   HelpfulnessNumerator    568454 non-null  int64
 5   HelpfulnessDenominator  568454 non-null  int64
 6   Score                   568454 non-null  int64
 7   Time                    568454 non-null  int64
 8   Summary                 568427 non-null  object
 9   Text                    568454 non-null  object
dtypes: int64(5), object(5)
memory usage: 43.4+ MB
```

FIGURE 2.2

```
Nombre de colonnes vides : 0
```

FIGURE 2.3

Les éléments vides et leur suppression.

Résultats de la suppression.

nombre de doublons

```
Nombre de lignes vides : 0
```

FIGURE 2.4

```
Id                      0
ProductId               0
UserId                  0
ProfileName            26
HelpfulnessNumerator    0
HelpfulnessDenominator  0
Score                   0
Time                    0
Summary                27
Text                    0
```

FIGURE 2.5

```
Id                      0
ProductId               0
UserId                  0
ProfileName             0
HelpfulnessNumerator    0
HelpfulnessDenominator  0
Score                   0
Time                    0
Summary                 0
Text                    0
dtype: int64
```

FIGURE 2.6

```
Nombre de lignes dupliquées : 0
```

FIGURE 2.7

convirtir time en s

```
0         1303862400
1         1346976000
2         1219017600
3         1307923200
4         1350777600
             ...
568449    1299628800
568450    1331251200
568451    1329782400
568452    1331596800
568453    1338422400
Name: Time, Length: 568401, dtype: int64
```

FIGURE 2.8

voila le resulta

```
0        2011-04-27
1        2012-09-07
2        2008-08-18
3        2011-06-13
4        2012-10-21
            ...
568449   2011-03-09
568450   2012-03-09
568451   2012-02-21
568452   2012-03-13
568453   2012-05-31
Name: Time, Length: 568454, dtype: datetime64[ns]
```

FIGURE 2.9

2.4 visualisation

La visualisation des données est un outil important pour analyser les données et communiquer des informations. C'est un moyen de transformer une grande quantité de données dans un format visuel, afin qu'elles puissent être facilement interprétées par les utilisateurs Et pour cela, nous faisons quelques visualisations.

Ce graphe contient le nombre de personnes ayant donné un score jusqu'à cinq.

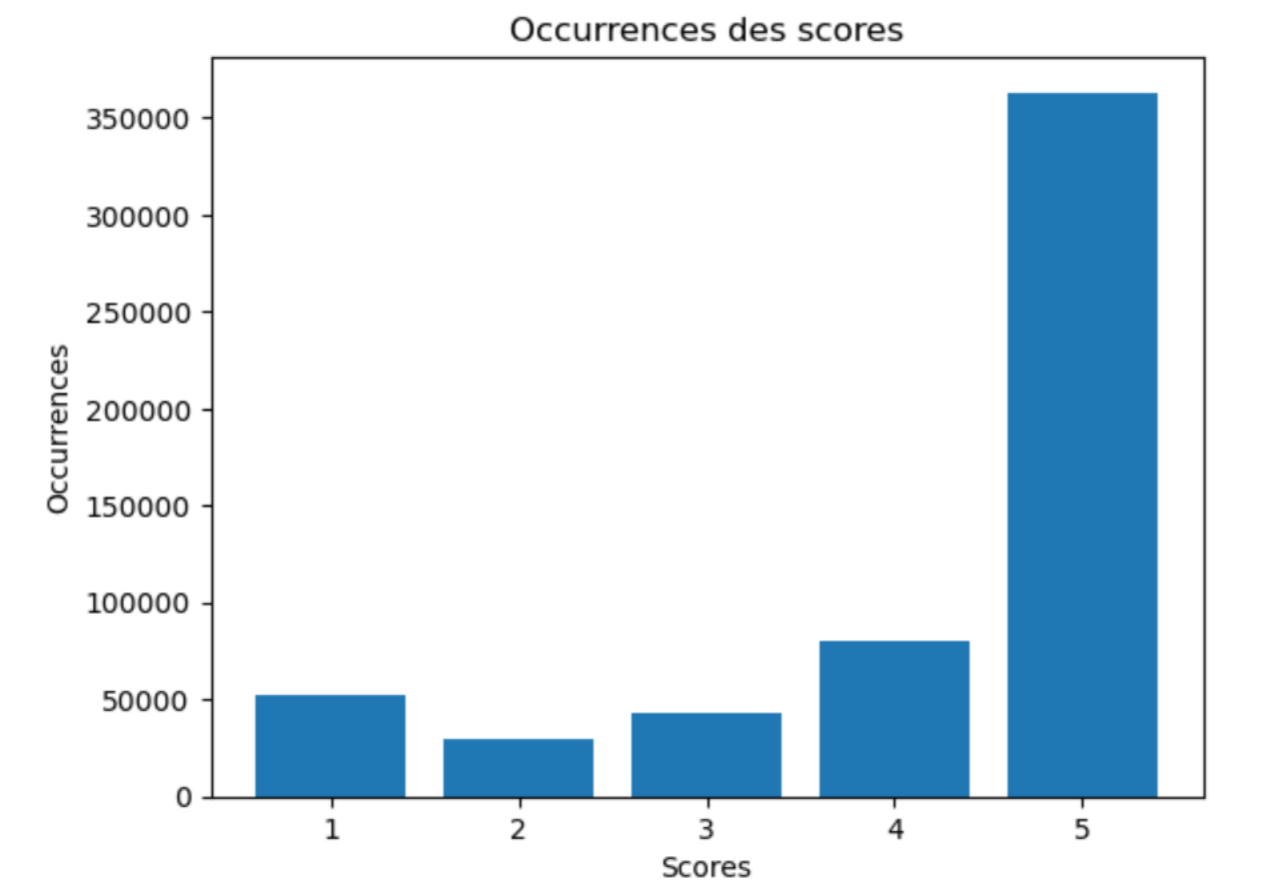


FIGURE 2.10

2.5 modilisation

Correction : Pour cette partie, j'ai utilisé deux modèles, Vader et Decision Tree, afin de faire une comparaison et de choisir le modèle le plus fiable et performant dans ses résultats. Avant de procéder à cela, nous avons effectué un prétraitement du texte, incluant la tokenisation et la vectorisation.

VADER (Valence Aware Dictionary and sEntiment Reasoner) est un outil d'analyse de sentiment basé sur un lexique et des règles, conçu spécifiquement pour détecter les sentiments

exprimés dans les médias sociaux. Il est entièrement open source sous la [Licence MIT] (nous apprécions sincèrement toutes les attributions et acceptons volontiers la plupart des contributions, mais veuillez ne pas nous tenir responsables).

text avant traitement :

```
print(df['Text'])
```

```
0         I have bought several of the Vitality canned d...
1         Product arrived labeled as Jumbo Salted Peanut...
2         This is a confection that has been around a fe...
3         If you are looking for the secret ingredient i...
4         Great taffy at a great price.  There was a wid...
                              ...
568449    Great for sesame chicken..this is a good if no...
568450    I'm disappointed with the flavor. The chocolat...
568451    These stars are small, so you can give 10-15 o...
568452    These are the BEST treats for training and rew...
568453    I am very satisfied ,product is as advertised,...
Name: Text, Length: 568454, dtype: object
```

FIGURE 2.11

text apres traitement :

```
0         [bought, several, vitality, canned, dog, food,...
1         [product, arrived, labeled, jumbo, salted, pea...
2         [confection, around, centuries, light, pillowy...
3         [looking, secret, ingredient, robitussin, beli...
4         [great, taffy, great, price, wide, assortment,...
                              ...
568449    [great, sesame, chickenthis, good, better, res...
568450    [im, disappointed, flavor, chocolate, notes, e...
568451    [stars, small, give, 1015, one, training, sess...
568452    [best, treats, training, rewarding, dog, good,...
568453    [satisfied, product, advertised, use, cereal, ...
Name: Processed_text, Length: 568401, dtype: object
```

FIGURE 2.12

```
Les 20 mots les plus fréquents :
 [('br', 264688), ('like', 251863), ('good', 195335), ('one', 172302), ('taste', 166572), ('g
55), ('product', 146438), ('flavor', 142441), ('tea', 133094), ('love', 126635), ('food', 123
t', 108169), ('really', 100413), ('dont', 95555), ('much', 91906), ('also', 86099), ('little'
```

FIGURE 2.13

Ce graphique représente les mots les plus fréquents dans le texte, en mettant en évidence particulièrement les mots associés à une connotation positive.

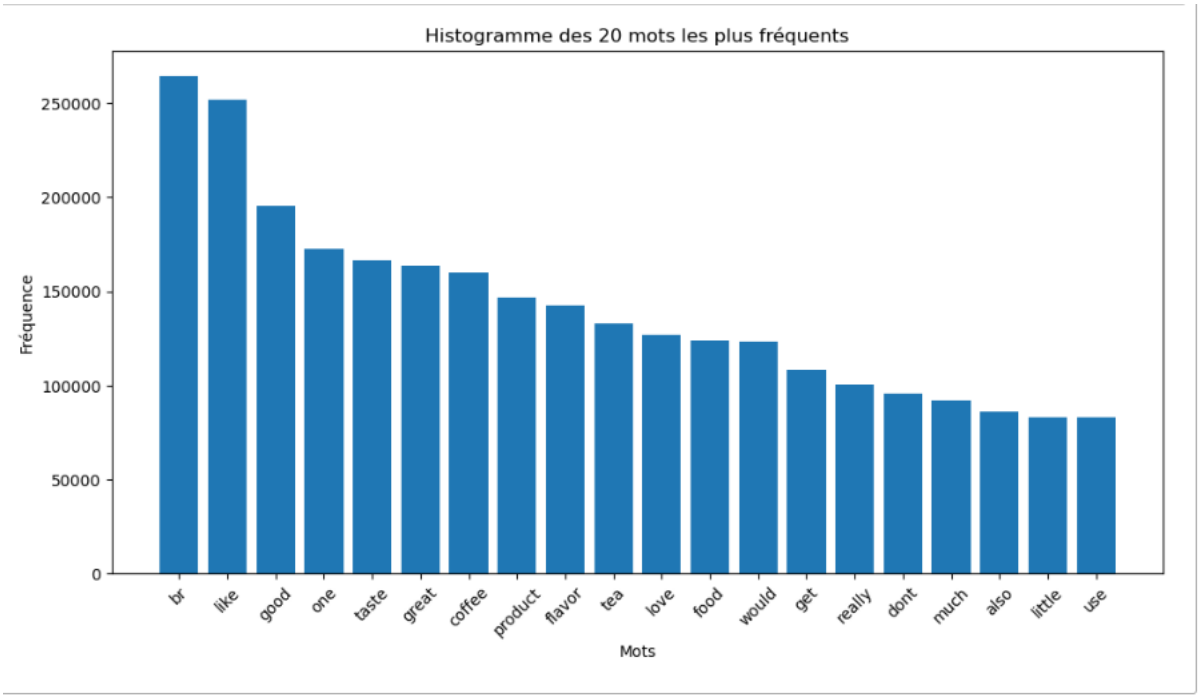


FIGURE 2.14

Le modèle ne comprend que des zéros, donc nous avons besoin de procéder à la vectorisation des chaî

```
Matrice de features vectorisées :
[[0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 ...
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]]

Shape de la matrice de features vectorisées : (568401, 1000)
```

FIGURE 2.15

2.6 Résultats du modèlel

Vader a produit des résultats à partir des scores, ensuite nous avons créé une autre table, effectué la jointure et visualisé les résultats. Ensuite, j'ai utilisé un autre modèle, Décision tree, pour comparer les résultats.

resultat vader :

```
0       {'neg': 0.0, 'neu': 0.695, 'pos': 0.305, 'comp...      5
1       {'neg': 0.138, 'neu': 0.862, 'pos': 0.0, 'comp...      1
2       {'neg': 0.091, 'neu': 0.754, 'pos': 0.155, 'co...      4
3       {'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...      2
4       {'neg': 0.0, 'neu': 0.552, 'pos': 0.448, 'comp...      5
...                                                          ...
568449  {'neg': 0.072, 'neu': 0.6, 'pos': 0.327, 'comp...      5
568450  {'neg': 0.19, 'neu': 0.697, 'pos': 0.114, 'com...      2
568451  {'neg': 0.037, 'neu': 0.884, 'pos': 0.078, 'co...      5
568452  {'neg': 0.041, 'neu': 0.506, 'pos': 0.452, 'co...      5
568453  {'neg': 0.0, 'neu': 0.846, 'pos': 0.154, 'comp...      5
```

FIGURE 2.16

apres la jointure : [1-3]

Text	SentimentScores	Label	PredictedLabel	Negative	Neutral	Positive	Compound	Id	ProductId	UserId	ProfileName	Help
I have bought several of the Vitality canned d...	{'neg': 0.0, 'neu': 0.695, 'pos': 0.305, 'comp...	5	positive	0.000	0.695	0.305	0.9441	1	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	
Product arrived labeled as Jumbo Salted Peanut...	{'neg': 0.138, 'neu': 0.862, 'pos': 0.0, 'comp...	1	negative	0.138	0.862	0.000	-0.5664	2	B00813GRG4	A1D87F6ZCVE5NK	dll pa	
This is a											Natalia	

FIGURE 2.17

la visualisation des resultats es comme suites :

1 tracer les resulta de vader :

En utilisant les scores que vous avez fournis (1, 2, 3, 4, et 5), avec les occurrences respectives (11.02%,4.93%, 7.11%,11.02% , et 63.23%), on peut voir une concentration significative de scores élevés (5) comparé aux autres scores. Cela pourrait indiquer que la plupart des évaluations ou opinions sont très positives.

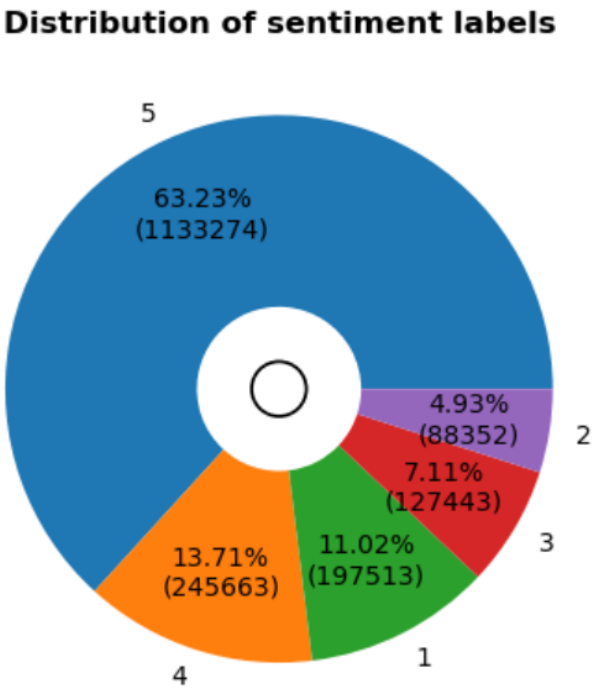


FIGURE 2.18

voila le graphe de chaques score : pour compound :

Les évaluations montrent une tendance très positive, avec une augmentation constante du sentiment positif à mesure que les scores augmentent. Les commentaires associés au score 5 obtiennent un 'compound score' particulièrement élevé, indiquant une satisfaction générale exceptionnelle.

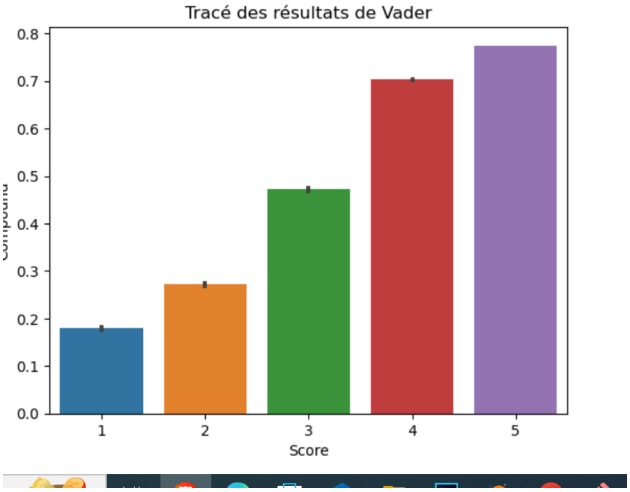


FIGURE 2.19

Tendance Négative :

Les évaluations révèlent une tendance marquée vers le négatif, avec une détérioration constante du sentiment à mesure que les scores diminuent.

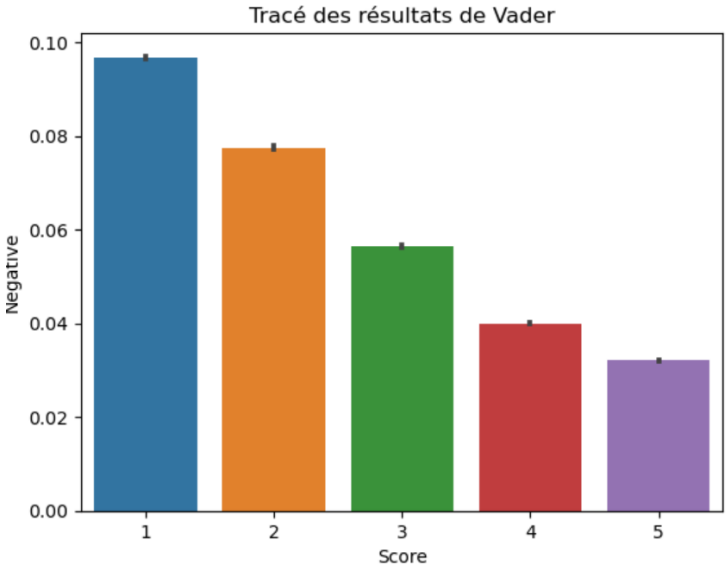


FIGURE 2.20

tendence posifive : Les évaluations dépeignent une tendance extrêmement positive, avec un accroissement régulier du sentiment positif à mesure que les scores augmentent. Les commentaires liés au score 5

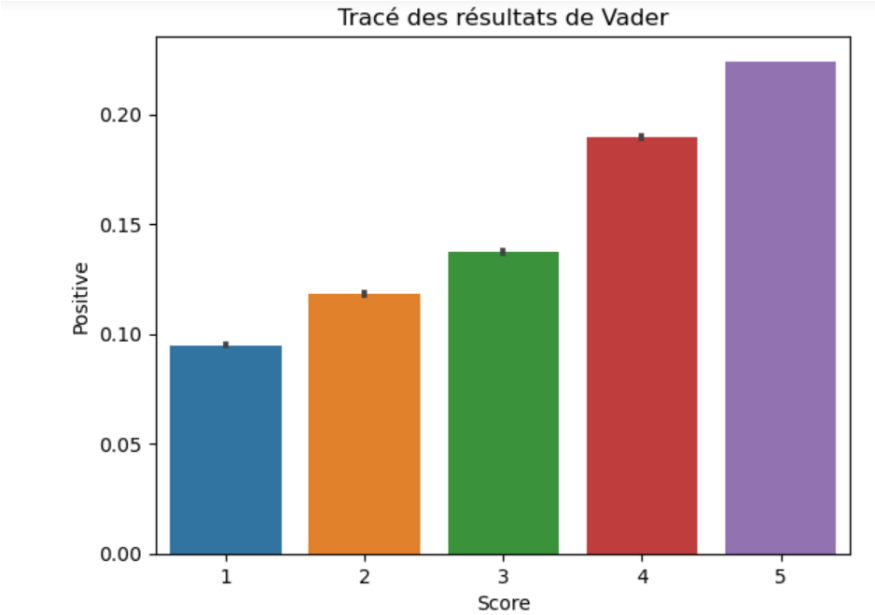


FIGURE 2.21

Tendance Neutre :

Les évaluations montrent une distribution relativement équilibrée, avec des ' scores' qui ne montrent pas de tendance claire vers la positivité ou la négativité.

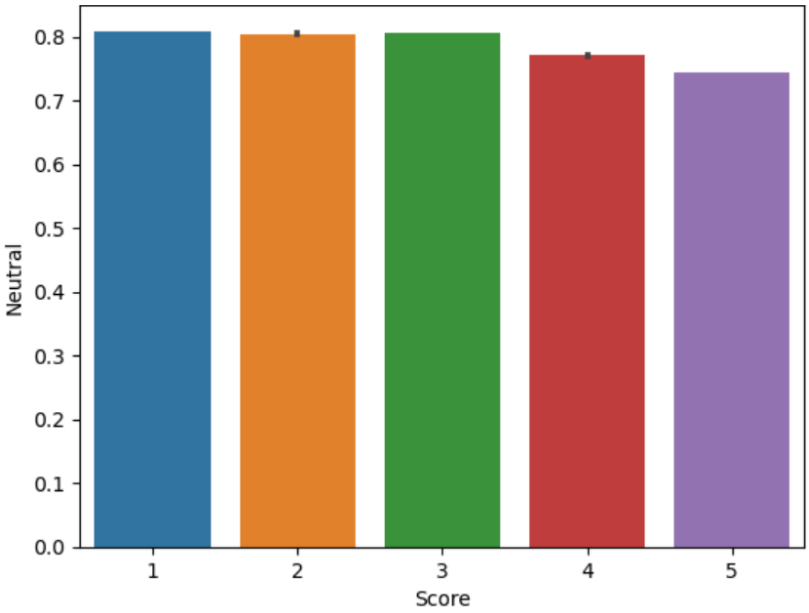


FIGURE 2.22

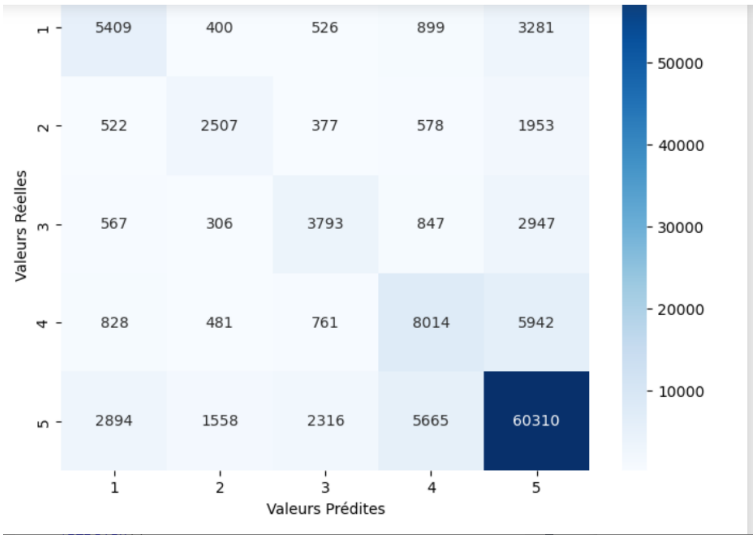
pour decision tree ;

```
Training Decision Tree Classifier...
Decision Tree Accuracy: 0.7040138633544744
Decision Tree Classification Report:
              precision    recall  f1-score   support

           1       0.53      0.51      0.52     10515
           2       0.48      0.42      0.45      5937
           3       0.49      0.45      0.47      8460
           4       0.50      0.50      0.50     16026
           5       0.81      0.83      0.82     72743

    accuracy                           0.70    113681
   macro avg       0.56      0.54      0.55    113681
weighted avg       0.70      0.70      0.70    113681

================================================
```

FIGURE 2.23

La matrice de confusion est une représentation visuelle des performances d'un modèle de classification. Elle permet de voir le nombre de vrais positifs, de vrais négatifs, de faux positifs et de faux négatifs. Dans le code, la matrice est affichée sous forme de heatmap. Cela offre une vue claire des prédictions du modèle par rapport aux valeurs réelles.

Les éléments diagonaux (TP, TN) représentent les prédictions correctes.
Les éléments hors diagonale (FP, FN) représentent les erreurs de prédiction



2.7 Interprétation des résultats

La comparaison entre le modèle Decision Tree et l'outil VADER pour l'analyse de sentiments révèle des différences notables dans leurs performances. Le modèle Decision Tree présente une précision globale de 70,4%, avec une focalisation particulière sur la classe de sentiments positifs, où il atteint une précision élevée de 81%. En revanche, l'outil VADER, basé sur une analyse des sentiments préconstruite, peut démontrer une approche plus rapide et souvent efficace, bien que moins personnalisable. Il est crucial de considérer les spécificités du contexte et les objectifs de l'analyse de sentiments pour choisir entre un modèle d'apprentissage automatique comme le Decision Tree, qui nécessite un entraînement sur des données spécifiques, et une solution prête à l'emploi comme VADER. L'évaluation continue des résultats et des ajustements peut aider à déterminer la meilleure approche en fonction des besoins spécifiques du projet.

2.8 Languaage utilisees

les languages utiliser dans ce projet :

python : ce language permet de lire et clairer data aussi la visualisation et pretretement des text sans oublier la musure de niveaux de satisfaction mon objecfif dans ce projet

Un excellent langage de programmation pour les projets NLP, notamment sa syntaxe simple et sa sémantique transparente. Les développeurs peuvent également accéder à d'excellents canaux de support pour l'intégration avec d'autres langages et outils

2.9 LES LOGICIELES :

— **Jupyter** : est une application web utilisée pour programmer dans plus de 40 langages de programmation, dont Python, Julia, Ruby, R, ou encore Scala.

CONCLUSION

En conclusion de ce projet axé sur le traitement du langage naturel (NLP) avec la modélisation en Python et l'apprentissage en profondeur, je suis extrêmement satisfait(e) d'avoir acquis une nouvelle expérience significative dans ma vie. Ce projet m'a permis de plonger profondément dans le domaine complexe du NLP, en utilisant des outils tels que la modélisation en Python et les techniques d'apprentissage en profondeur.

L'étude approfondie des modèles tels que Vader et l'Arbre de Décision m'a offert une compréhension approfondie des mécanismes sous-jacents du NLP, tout en me confrontant aux défis concrets liés au nettoyage des données, à la vectorisation et à l'analyse des sentiments. Cette expérience m'a également permis d'affiner mes compétences en Python, renforçant ainsi ma capacité à implémenter des solutions pratiques pour des problématiques complexes.

Bibliographie

[1] https ://fr.oncrawl.com/seo-technique/utiliser-le-nlp-pour-ameliorer-lexperience-client/

[2] https ://www.kaggle.com/datasets/snap/amazon-fine-food-reviews/data/

[3] https ://vadersentiment.readthedocs. io/en/latest/

[4] https ://www.tableau.com/learn/articles/what is datacleaning : :text=Data