

RÉALISÉ PAR :
GAJJA NOUR EDDIN

ENCADRÉ PAR :
Mr CHARAF HAMIDI

MEMBRES DE JURY :
Mr CHARAF HAMIDI
Mlle SALMA GOAU

1 Introduction

2 OBJECTIFS DU PROJET

3 Méthodologie

4 modèles et ces résultats

5 CONCLUSION

INTRODUCTION ET CONTEXTE

NLP signifie Traitement du Langage Naturel. C'est un domaine de l'intelligence artificielle qui se concentre sur la capacité des ordinateurs à comprendre, interpréter et générer le langage humain de manière significative et contextuellement pertinente.

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,

1 Introduction

2 OBJECTIFS DU PROJET

3 Méthodologie

4 modèles et ces résultats

5 CONCLUSION

OBJECTIFS

Ce projet vise à utiliser 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 Introduction

2 OBJECTIFS DU PROJET

3 Méthodologie

4 modèles et ces résultats

5 CONCLUSION

les étapes de réalisation du projet.

1 Téléchargement des données et leur nettoyage:

```
In [2]: df = pd.read_csv('C:\\Users\\HP\\Desktop\\dataproyet\\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 | dill 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... |

Figure 1:

2 la description de base de donner :

Dans cette base de données, nous avons dix colonnes différentes et 568 454 lignes. Nous avons besoin d'effectuer un nettoyage des données


```
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:

3 nettoyage des données :

Avant de faire la visualisation pour mieux comprendre la base de données et la relation entre les colonnes, il est important d'effectuer un nettoyage des données. Voici quelques captures d'écran:

Voici le nombre des colonnes vides.

Nombre de colonnes vides : 0

Figure 3:

Voici les nombre des lignes vides.

Nombre de lignes vides : 0

Figure 4:

Les éléments vides et leur suppression.

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

Figure 5:

Résultats de la suppression.

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

Figure 6:

nombre de doublons

Nombre de lignes dupliquées : 0

Figure 7:

convertir 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 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 9:

(3-1) prétraitement de texte :

Le prétraitement de texte est une étape cruciale dans le domaine du traitement naturel du langage (NLP)

Conversion en minuscules (lowercasing) : Uniformisez la casse des mots en les convertissant tous en minuscules. Cela évite les variations dues à la casse et facilite la comparaison des mots.

Tokenisation : Divisez le texte en unités plus petites appelées "tokens". Un token peut être un mot ou une sous-chaîne de mots. Cela rend le texte plus maniable pour l'analyse.

Suppression des mots vides (stopwords) : Éliminez les mots courants qui n'apportent pas beaucoup de sens à l'analyse, comme "le", "la", "de", etc.

Lemmatisation et racinisation : Réduisez les mots à leur forme de base. La lemmatisation ramène les mots à leur forme canonique, tandis que la racinisation réduit les mots à leur racine.

En effectuant ces étapes de manière efficace, on obtient ces résultats:
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, robituussin, 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 10:

- 1 Introduction
- 2 OBJECTIFS DU PROJET
- 3 Méthodologie
- 4 modèles et ces résultats**
- 5 CONCLUSION

J'ai utilisé deux modèles pour effectuer une analyse approfondie. Initialement, les scores générés par le modèle Vader ont été examinés pour évaluer les résultats obtenus. Par la suite, afin de procéder à une comparaison approfondie et d'explorer davantage les nuances du sentiment, j'ai également intégré un autre modèle, à savoir le Décision Tree. Cette approche combinée m'a permis d'obtenir des perspectives plus riches et de mieux comprendre les variations dans l'analyse de sentiment effectuée par ces deux modèles distincts.

resultat vader :

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

Figure 11:

la visualisation des resultats es comme suites :

1 tracer les resulta de vader :

Distribution of sentiment labels

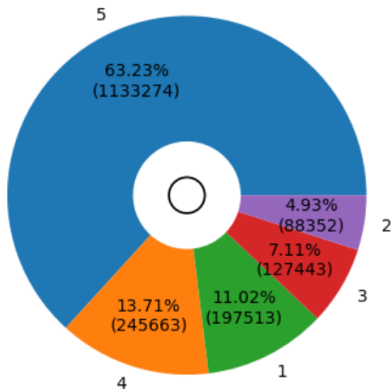


Figure 12:

pour decision tree ;

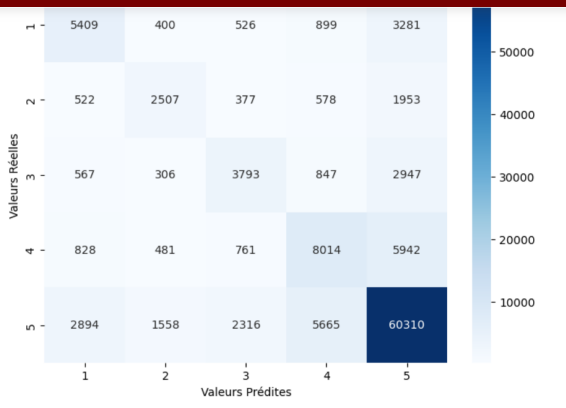
```
Training Decision Tree Classifier...
Decision Tree Accuracy: 0.7040138633544744
Decision Tree Classification Report:
              precision    recall  f1-score   support

         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 13:



elements diagonaux (TP, TN)representent lespredictions correctes. Les el ements hors diagonale (FP, FN) representent les erreurs de predic-tion.

1 Introduction

2 OBJECTIFS DU PROJET

3 Méthodologie

4 modèles et ces résultats

5 CONCLUSION

Ce projet NLP en Python m'a offert une expérience enrichissante, explorant des modèles tels que Vader et l'Arbre de Décision. En plongeant dans le domaine complexe du NLP, j'ai renforcé mes compétences en Python et perfectionné mes capacités pour résoudre des problématiques complexes liées au nettoyage des données, à la vectorisation et à l'analyse des sentiments.