الكلية متعددة التخصصات-ورزازات
ⵜⴰⴽⵍⵉⵢⵜ ⵜⴰⵅⵜⵚⵚⵉⵜ – ⵍⵓⵕⵣⴰⵣⴰⵜ
FACULTÉ POLYDISCIPLINAIRE DE OUARZAZATE

L'analyse des sentiments avec NLP

BOUHLALI Abdelfattah

Master Mathématiques Appliquées pour la Science des Données

2023 - 2024

Encadré par :
GAOU Salma
HAMIDI Charaf

Résumé

Le projet fait partie d'une enquête approfondie visant à décrypter les sentiments exprimés dans les avis des consommateurs sur la plateforme Amazon, en se concentrant spécifiquement sur les produits alimentaires. Notre principal objectif était d'utiliser des techniques avancées de traitement du langage naturel (NLP) et des modèles pré-entraînés, tels que RoBERTa, pour analyser et interpréter les opinions des utilisateurs, afin d'identifier des tendances significatives et de mettre en lumière les sentiments prédominants au sein de cette catégorie de produits.

Summary :

The project is part of a comprehensive investigation designed to decipher the sentiments expressed in consumer reviews on the Amazon platform, with a specific focus on food products. Our main objective was to utilize advanced natural language processing (NLP) techniques and pre-trained models, such as RoBERTa, to analyze and interpret user opinions, in order to identify significant trends and highlight prevailing sentiments within this product category.

Table des matières

1 Introduction

Notre projet explore les avis des consommateurs sur Amazon, en se concentrant sur les produits alimentaires. Le défi est de comprendre ce que les clients pensent vraiment. Avec tant d'avis, il est difficile de trouver les informations importantes. Nous voulons transformer ces avis en idées utiles pour aider les entreprises à améliorer leurs produits et à satisfaire les clients.

1.1 Problème ou Question à Résoudre

On essaie de comprendre les avis des gens sur les produits alimentaires d'Amazon. Comment les clients se sentent-ils vraiment ? C'est difficile car il y a beaucoup d'avis. Notre but est de trouver des informations importantes pour aider les entreprises.

1.2 Contexte et Motivation

Beaucoup de gens achètent sur Amazon, et ils laissent beaucoup d'avis. Mais ces avis ne sont pas toujours faciles à comprendre. Nous voulons aider les entreprises à comprendre ce que les clients aiment et n'aiment pas.

1.3 Objectifs du Projet et Hypothèses à Tester

Pour notre projet, nous avons défini plusieurs objectifs clés :

1. **Comprendre les Sentiments :** Utiliser des outils avancés comme RoBERTa pour analyser les sentiments exprimés dans les avis sur les produits alimentaires d'Amazon, en identifiant s'ils sont positifs, négatifs ou neutres.
2. **Identifier les Tendances :** Découvrir les tendances émergentes dans les avis, y compris les préférences alimentaires, les aspects spécifiques appréciés ou critiqués, et les évolutions au fil du temps.
3. **Améliorer la Pertinence :** Déterminer les aspects les plus importants pour les clients en analysant les mots clés et les expressions fréquemment utilisés dans les avis.

Hypothèses à Tester :

1. Nous supposons que les sentiments des clients varient en fonction des types de produits alimentaires, et nous chercherons à identifier ces variations.
2. Nous pensons que certains mots-clés auront une influence significative sur la perception des produits, et nous testerons cette hypothèse en analysant leur fréquence.
3. Nous anticipons que les tendances dans les avis sur les produits alimentaires évoluent avec le temps, et nous chercherons à confirmer cette hypothèse en examinant les changements au fil des mois.

2 Méthodologie

2.1 Données

2.1.1 Source des Données

Les données utilisées dans cette étude proviennent d'un ensemble de critiques sur des produits alimentaires provenant d'Amazon. Ce jeu de données couvre une période de plus de 10 ans, comprenant l'ensemble des 500,000 critiques jusqu'en octobre 2012. Les critiques incluent des informations sur les produits et les utilisateurs, les évaluations et une critique en texte brut. De plus, il englobe des critiques de toutes les autres catégories d'Amazon.

Le lien vers les données est disponible sur Kaggle : `https://www.kaggle.com/snap/amazon-fine-food-reviews`.

2.1.2 Les composants d'un avis

FIGURE 1 – Les elements d'un avis

Un avis typique de cet ensemble de données comprend plusieurs composants essentiels qui fournissent une perspective détaillée sur l'expérience de l'utilisateur. Voici une énumération des principaux éléments d'un avis dans cet ensemble de données :

1. **Identifiant Unique ('Id') :** Chaque avis est associé à un identifiant unique, permettant une référence spécifique.

2. **Code Produit ('ProductId') :** Indique le produit concerné par l'avis, facilitant l'association aux références produits.

3. **Identifiant de l'Utilisateur ('UserId') :** Identifie de manière unique l'utilisateur ayant publié l'avis.

4. **Nom du Profil de l'Utilisateur ('ProfileName') :** Le nom du profil associé à l'utilisateur, fournissant un contexte sur l'auteur de l'avis.

5. **Utilité ('HelpfulnessNumerator' et 'HelpfulnessDenominator') :** Deux valeurs numériques indiquant le nombre d'utilisateurs qui ont trouvé l'avis utile par rapport au nombre total d'évaluations de son utilité.

6. **Notation ('Score') :** La notation attribuée par l'utilisateur, quantifiant l'appréciation globale du produit.

7. **Timestamp de l'Avis ('Time') :** La date et l'heure à laquelle l'avis a été publié, offrant une dimension temporelle.

8. **Résumé ('Summary') :** Un condensé du contenu de l'avis, fournissant une vue d'ensemble rapide.

9. **Texte Intégral de l'Avis ('Text') :** Le contenu complet de l'avis, offrant des détails contextuels sur l'expérience de l'utilisateur.

Chacun de ces éléments joue un rôle spécifique dans la caractérisation de l'avis, permettant une analyse approfondie des sentiments exprimés dans les avis sur les produits alimentaires de la plateforme Amazon.

2.1.3 Informations générales sur la dataset

```
print(data.info())
```

```
<class 'pandas.core.frame.DataFrame'>
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
None
```

FIGURE 2 – Les information sur la DataSet

La sortie de la fonction `data.info()` fournit une vue détaillée de la structure de notre ensemble de données. Notre ensemble de données est représenté sous la forme d'un objet de type `pandas.core.frame.DataFrame`, avec un index de type `RangeIndex`, allant de 0 à 568453, indiquant ainsi le nombre total d'entrées (lignes) dans la base de données. Il est composé de 10 colonnes au total.

Les attributs et types de données de chaque colonne sont les suivants :

— `'Id'` est de type `int64` avec 568454 valeurs non nulles.
— `'ProductId'` est de type `object` (généralement une chaîne de caractères) avec 568454 valeurs non nulles.
— `'UserId'` est de type `object` avec 568454 valeurs non nulles.
— `'ProfileName'` est de type `object` avec 568428 valeurs non nulles, et présente 26 valeurs manquantes.
— `'HelpfulnessNumerator'` est de type `int64` avec 568454 valeurs non nulles.
— `'HelpfulnessDenominator'` est de type `int64` avec 568454 valeurs non nulles.
— `'Score'` est de type `int64` avec 568454 valeurs non nulles.
— `'Time'` est de type `int64` avec 568454 valeurs non nulles.
— `'Summary'` est de type `object` avec 568427 valeurs non nulles, mais présente 27 valeurs manquantes.
— `'Text'` est de type `object` avec 568454 valeurs non nulles.

La mémoire utilisée par cet ensemble de données est d'environ 43.4 MB. Il est également pertinent de noter que `'ProfileName'` a 26 valeurs manquantes, tandis que `'Summary'` a 27 valeurs manquantes. Ces informations sont essentielles pour appréhender la composition de notre ensemble de données, notamment en termes de types de données, de présence de valeurs manquantes, et de la mémoire occupée par l'ensemble de données.

2.1.4 Statistiques descriptives pour les attributs numériques

```
print(data.info())
```

```
<class 'pandas.core.frame.DataFrame'>
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
None
```

FIGURE 3 – Les information sur la DataSet

2.1.5 Traitement des Données

Avant d'entamer l'analyse, un processus de nettoyage des données a été appliqué. Cela inclut la suppression des données manquantes, la normalisation des textes, et la gestion des éventuels biais. De plus, une étape de pré-traitement a été effectuée pour optimiser la qualité des données avant l'application des modèles.

2.2 Techniques d'Analyse de Données et de Machine Learning

2.2.1 Traitement du Langage Naturel (NLP)

L'analyse des sentiments repose sur des techniques avancées de Traitement du Langage Naturel (NLP). Nous avons choisi d'utiliser le modèle pré-entraîné RoBERTa, reconnu pour sa performance dans la compréhension contextuelle des textes.

2.2.2 Algorithme d'Analyse des Sentiments

Pour notre étude, nous avons opté pour un algorithme d'analyse des sentiments basé sur l'apprentissage profond. Ce choix a été motivé par la capacité des réseaux neuronaux à capturer des relations complexes dans les données textuelles.

2.3 Justification des Méthodes et Algorithmes

Le choix du modèle RoBERTa a été guidé par sa réputation pour la compréhension fine des nuances linguistiques, essentielle dans l'analyse des sentiments. De plus, l'utilisation d'un modèle pré-entraîné permet de tirer parti de la richesse des données sur lesquelles il a été initialement formé.

L'algorithme d'analyse des sentiments basé sur l'apprentissage profond a été privilégié en raison de sa capacité à apprendre des représentations hiérarchiques complexes, améliorant ainsi la précision de la prédiction des sentiments.

2.4 Paramètres et Hyperparamétrage

Les paramètres du modèle RoBERTa ont été conservés conformément aux recommandations de l'auteur, avec une attention particulière portée à l'optimisation des performances. L'hyperparamétrage de l'algorithme d'analyse des sentiments a été ajusté de manière itérative pour maximiser la précision sans compromettre la généralisation.

Cette section a détaillé la méthodologie utilisée pour collecter, préparer et analyser les données, ainsi que les choix effectués en termes de techniques d'analyse de données et d'algorithmes. La justification de ces choix est cruciale pour garantir la fiabilité et la pertinence des résultats obtenus dans le cadre de cette étude.

3 Quelques commandes

Voici quelques commandes utiles :

3.1 Insertion de figures

Ici, je cite l'image 4 dans le texte.

3.2 Insertion d'équation

$$\rho + \Delta = 42 \tag{1}$$

L'équation 1 est citée ici.

3.3 Insertion d'une référence bibliographique

Les références (articles scientifiques, articles de journaux, blogs, pages web) doivent être mentionnées dans le texte par une balise [1] et fait le lien avec la citation incluse dans la bibliographie.

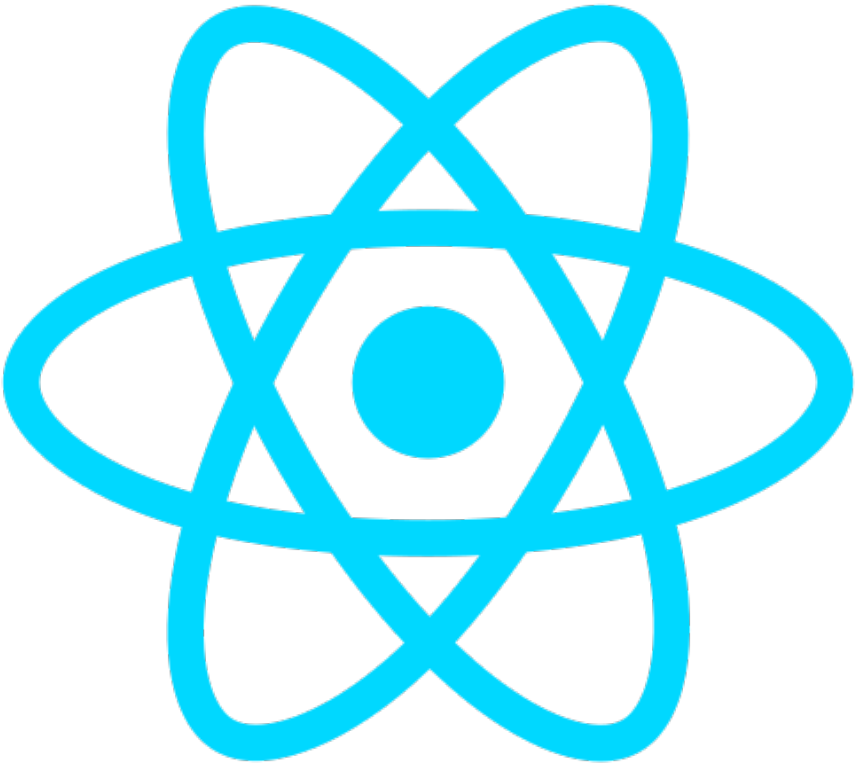

FIGURE 4 – Mettre une légende explicite à votre figure

Références

[1] Leslie Lamport, *LaTeX : A Document Preparation System*. Addison Wesley, Massachusetts, 2nd Edition, 1994.

Table des figures