Analyse des sentiments avec NLP

BOUHLALI Abdelfattah

Master Mathématiques Appliquées pour la Science des Données

2023 - 2024

الكلية متعددة التخصصات-ورزازات
FACULTÉ POLYDISCIPLINAIRE DE OUARZAZATE

Encadré par :
GAOU SALMA
HAMIDI CHARAF

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.

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.

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.

Objectifs du Projet et Hypothèses à Tester

Objectifs du Projet :

1. Comprendre les Sentiments.
2. Identifier les Tendances.
3. Améliorer la Pertinence.

Hypothèses à Tester :

1. Les sentiments des clients varient en fonction des types de produits alimentaires.
2. Certains mots-clés auront une influence significative sur la perception des produits.
3. Les tendances dans les avis sur les produits alimentaires évoluent avec le temps.

Étapes de la Méthodologie

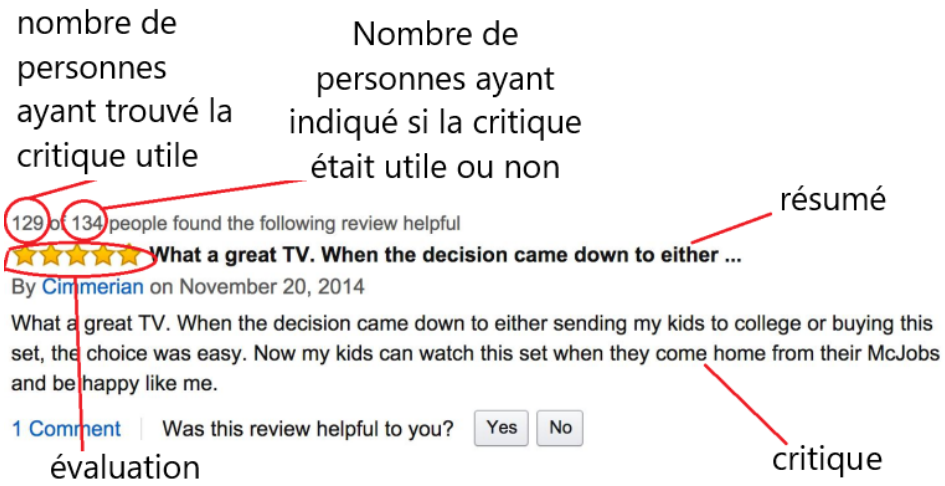

Figure 1: Les elements d'un avis

1 Collecte de Données :

- Collecte des avis sur les produits alimentaires d'Amazon.
- Utilisation de la fonction `data.info()` pour examiner la structure de l'ensemble de données.

2 Informations Générales sur la Dataset :

- Utilisation de la fonction `data.info()` pour obtenir une vue détaillée de la structure de l'ensemble de données.
- Analyse des attributs et types de données de chaque colonne.

```
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

Étapes de la Méthodologie (suite)

3. **Statistiques Descriptives pour les Attributs Numériques :**
   - Utilisation de la commande `data['Score'].describe()` pour obtenir des statistiques descriptives pour l'attribut "Score".

```
data['Score'].describe()
```

```
count    568454.000000
mean          4.183199
std           1.310436
min           1.000000
25%           4.000000
50%           5.000000
75%           5.000000
max           5.000000
Name: Score, dtype: float64
```

Figure 3: Statistiques descriptives pour l'attribut Score

4. **Vérification des Valeurs Manquantes ou d'Incohérences :**
   - Utilisation de la commande `data.isnull().sum()` pour détecter les valeurs manquantes dans chaque colonne.

Étapes de la Méthodologie (suite)

5. **Vérifier les Valeurs Uniques dans Chaque Colonne :**
   - Utilisation de la commande `data.nunique()` pour explorer la diversité des valeurs dans chaque colonne.
   - Interprétation des résultats.
6. **Prétraitement des Données :**
   - Élimination des lignes en double et gestion des valeurs manquantes avec `data.drop_duplicates()` et `data.dropna()`.

7 Analyse Exploratoire des Données :

- Distribution des scores.

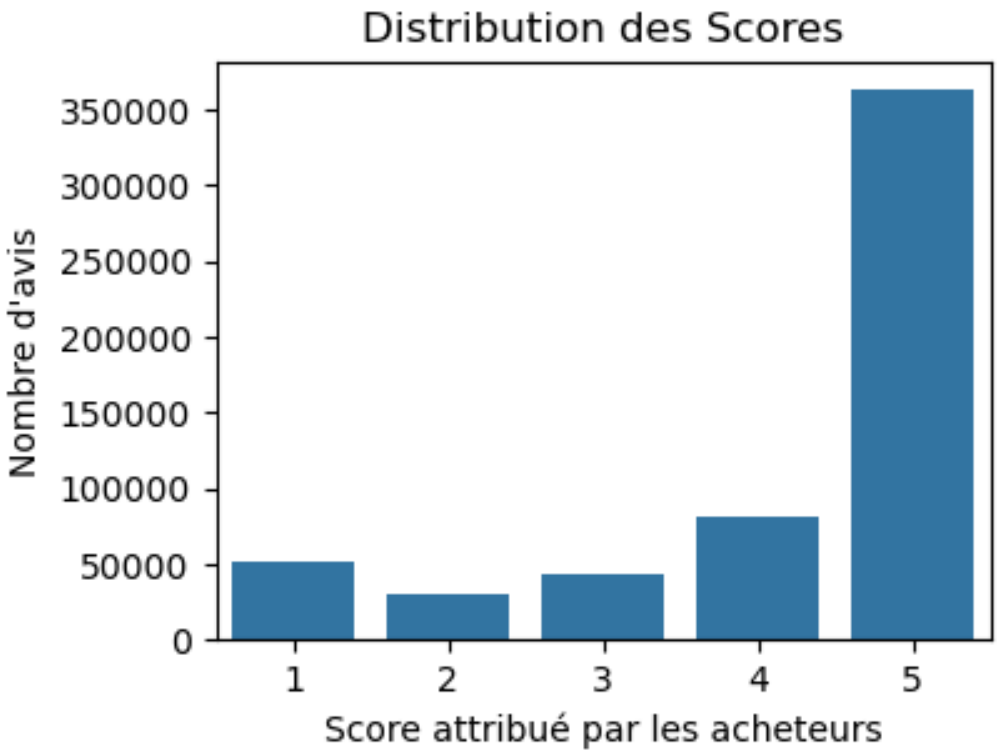

Figure 4: La distribution des scores

7 Analyse Exploratoire des Données :

- Calcul de la moyenne et de la médiane des scores.

Moyenne des scores : 4.18

Médiane des scores : 5.0

❼ Analyse Exploratoire des Données :

- Distribution des sentiments.

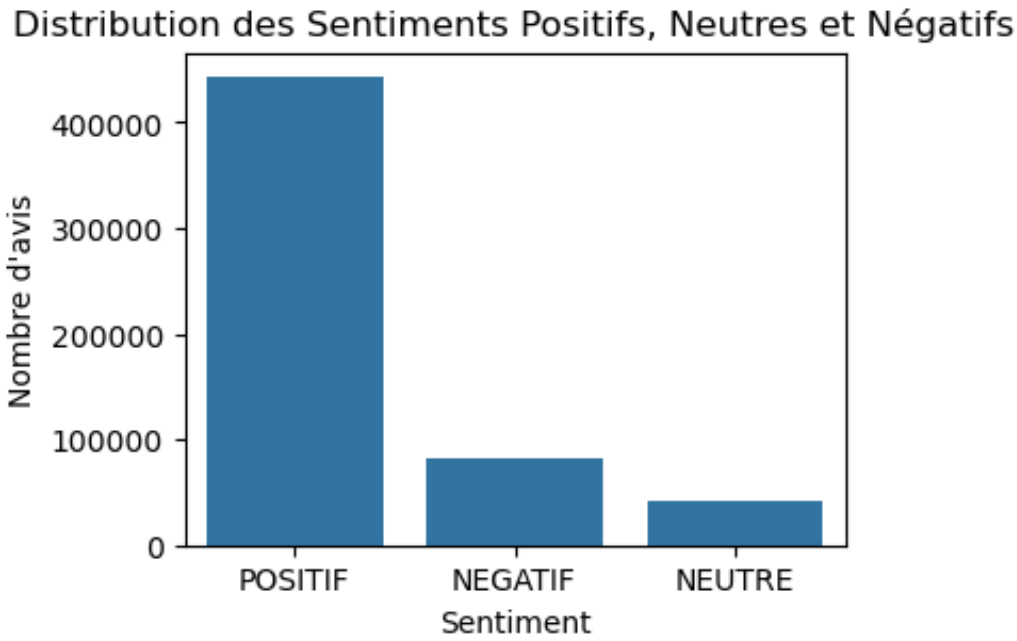

Figure 5: Distribution des Sentiments

Étapes de la Méthodologie (suite)

8. **Traitement du Texte** :
   - Suppression des URL, des balises HTML, des caractères non alphabétiques.
   - Conversion en minuscules et suppression des stopwords.
9. **Utilisation du Modèle Pré-entraîné RoBERTa** :
   - Initialisation du modèle RoBERTa et du tokenizer.
   - Fonction d'évaluation des scores de RoBERTa pour l'analyse de sentiments.
   - Analyse de sentiments avec RoBERTa sur l'ensemble de données.
10. **Transformation et Fusion des Résultats** :
    - Stockage des résultats du modèle avec les données d'origine.
    - Exportation des résultats en CSV (`nlp_results.csv`).

Analyse de la Répartition des Sentiments

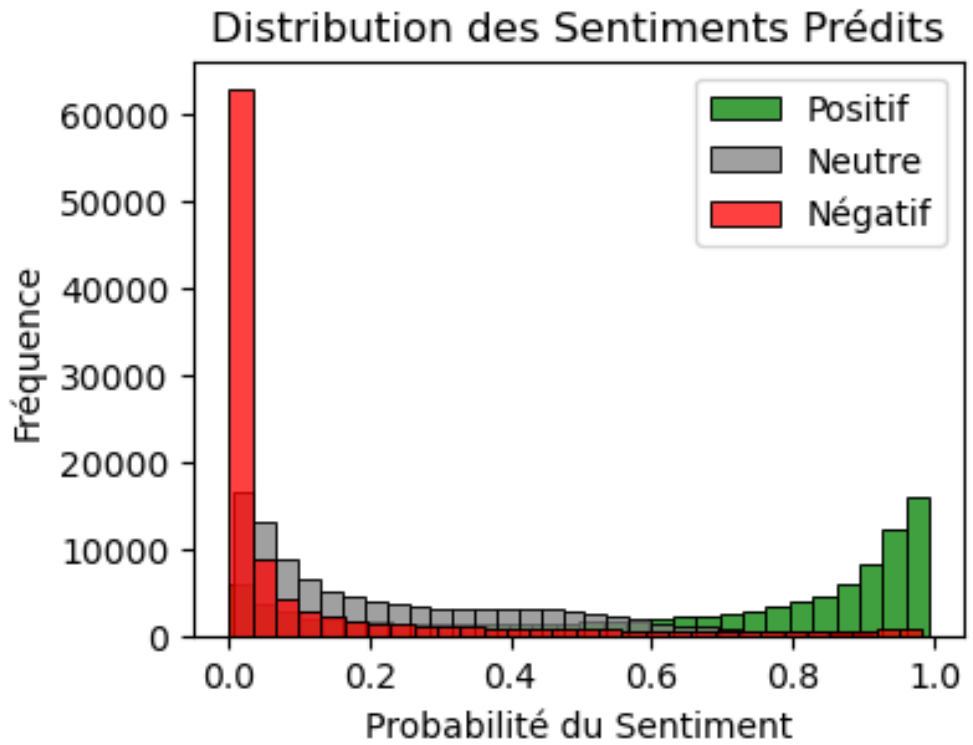

Figure 6: Distribution des Sentiments Prédits

L'histogramme des probabilités de sentiments prédites révèle des tendances distinctes dans la confiance du modèle.

Distribution des Sentiments Prédits par RoBERTa

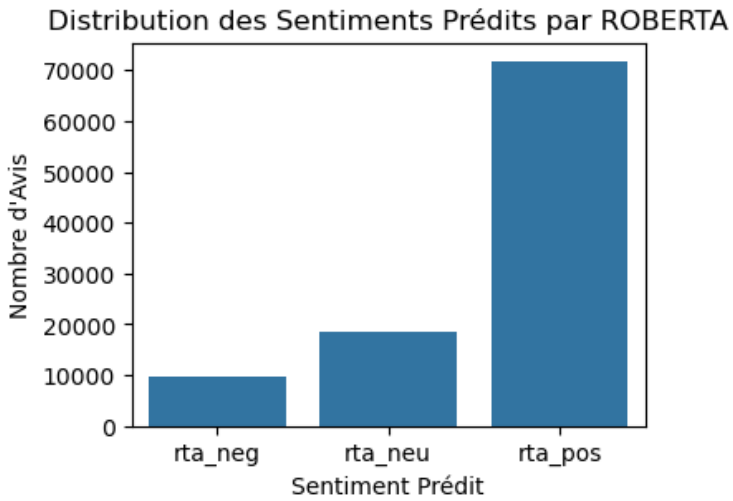

Figure 7: Distribution des Sentiments Prédits par RoBERTa

La répartition des sentiments prédits par RoBERTa sur l'ensemble de vos avis est la suivante :

- Sentiment Positif (rta_pos) : 71,827 occurrences
- Sentiment Neutre (rta_neu) : 18,492 occurrences
- Sentiment Négatif (rta_neg) : 9,607 occurrences

Distribution des Sentiments RoBERTa selon les Scores

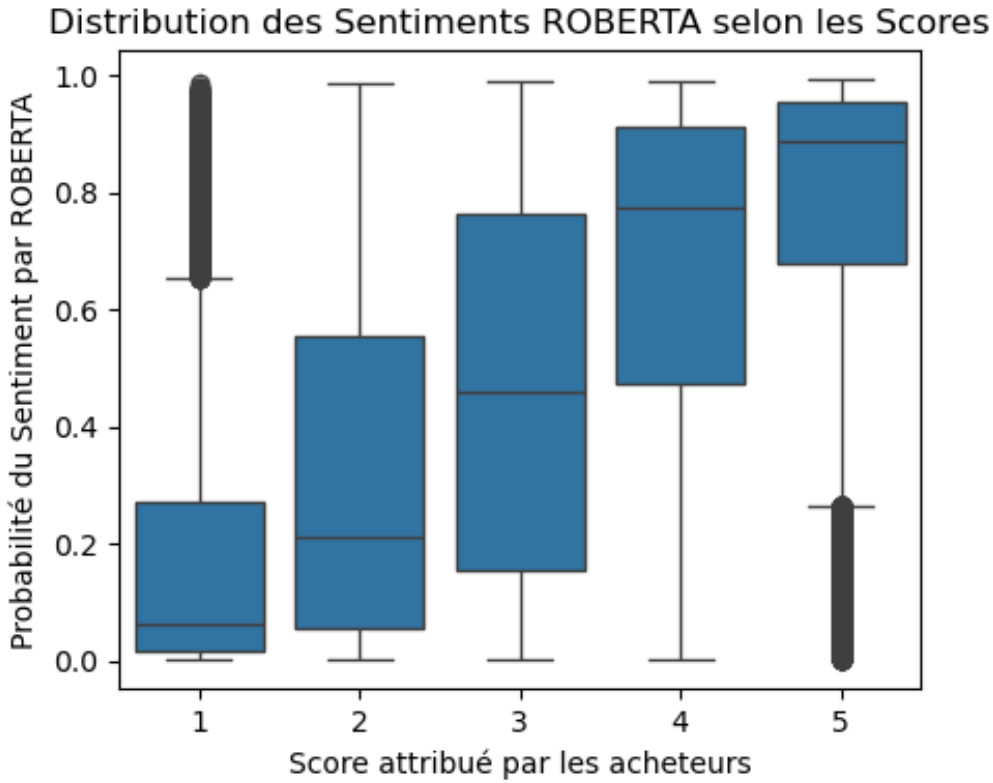

Figure 8: Distribution des Sentiments RoBERTa selon les Scores

On constate que la distribution des prédictions de sentiment générées par le modèle RoBERTA, classées selon les scores attribués par les acheteurs.

Diagramme circulaire des proportions de sentiments

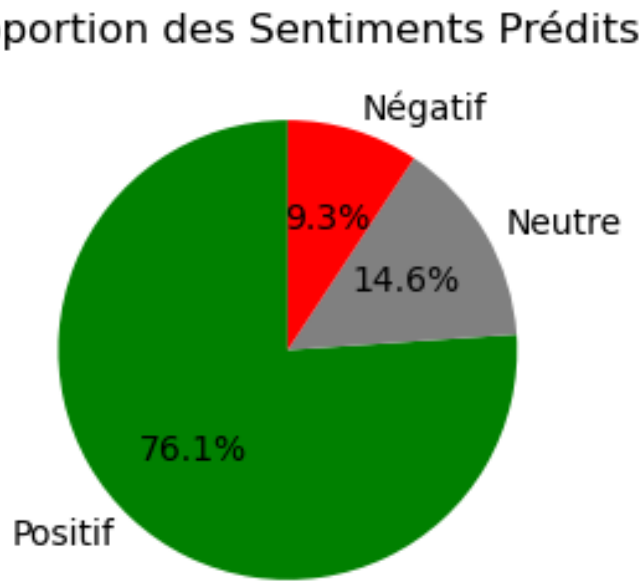

Figure 9: Diagramme circulaire des proportions de sentiments

L'analyse des prédictions de sentiments révèle une tendance marquée vers des sentiments positifs, avec une majorité écrasante de 76.1

Analyse des Relations entre les Probabilités Prédites et les Scores Utilisateur

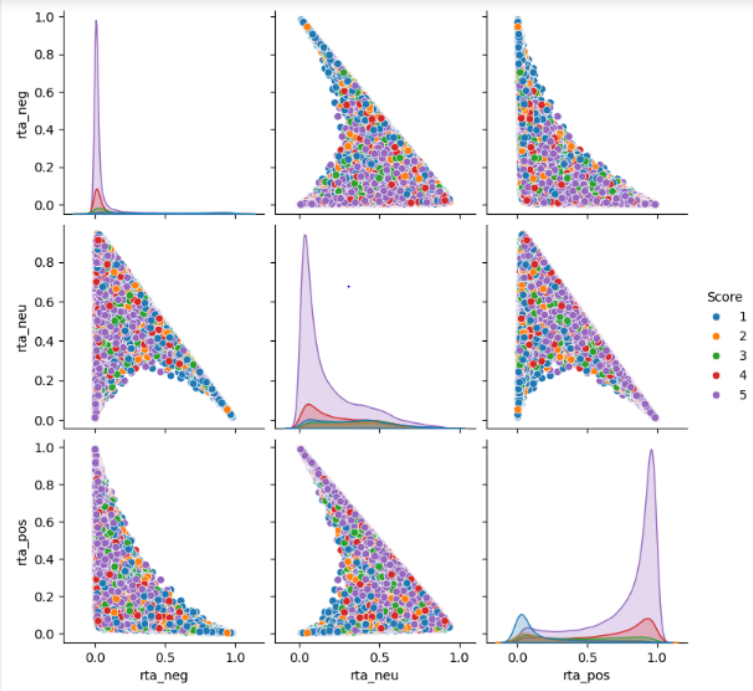

Figure 10: Relation entre les Probabilités Prédites et les Scores Utilisateur

L'analyse des résultats met en évidence la distribution des probabilités prédites pour chaque sentiment (négatif, neutre, positif) en fonction des scores attribués par les utilisateurs.

Nuage de mots des avis positifs

Figure 11: Nuage de Mots pour les Avis Positifs

Ce nuage de mots met en lumière les termes les plus fréquemment associés aux avis considérés comme positifs.

Nuage de mots des avis négatifs

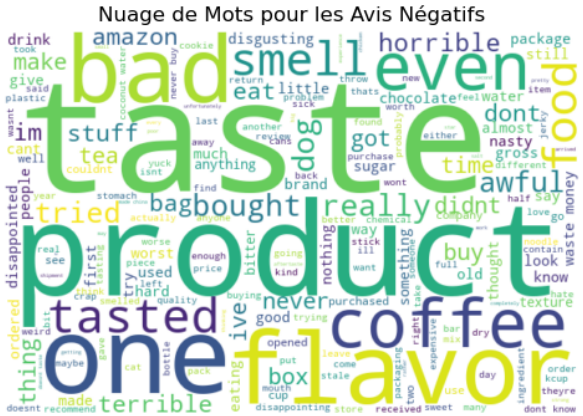

Figure 12: Nuage de Mots pour les Avis Négatifs

De manière similaire, le nuage de mots représente les termes les plus fréquemment associés aux avis considérés comme négatifs par le modèle RoBERTa.

Points forts de RoBERTa

- Compréhension fine des sentiments : RoBERTa excelle dans la compréhension fine des sentiments exprimés dans les avis, en distinguant clairement entre les sentiments positifs, négatifs et neutres.
- Capacité à identifier les tendances : RoBERTa peut découvrir les tendances émergentes dans les avis, y compris les préférences alimentaires, les aspects appréciés ou critiqués, offrant ainsi des informations précieuses sur les évolutions du marché.
- Analyse de mots clés et d'expressions fréquemment utilisés : RoBERTa peut aider à déterminer les aspects importants en analysant les mots clés et les expressions fréquemment utilisés dans les avis, permettant une identification rapide des points forts ou des points faibles des produits.

Limites de RoBERTa

- Dépendance aux données d'entraînement : RoBERTa dépend fortement des données sur lesquelles il a été formé, ce qui peut affecter sa précision.
- Interprétation des résultats : Les modèles NLP complexes comme RoBERTa peuvent être difficiles à interpréter, rendant délicate l'interprétation des résultats.
- Besoin de ressources informatiques importantes : RoBERTa nécessite des ressources informatiques importantes pour son utilisation.

Limites de RoBERTa

- Des Textes mal Classés : Malgré la plupart des classifications correctes, il peut y avoir des textes mal classés en raison de nuances linguistiques.

```
l_df.query("Score == 1").sort_values("rta_pos" , ascending=False)['Text'].values[0]
```

```
"I was excited to find this and read the great reviews.  I ordered it at $15+ from Amazon, and it does work and taste great
in our large school popcorn maker. HOWEVER, it's offered for $3.48 at our local food warehouse!!  Wow, Amazon is really maki
ng some money off this one.  Shop local before you buy this overpriced wonder."
```

ICI La critique est négatif. La personne exprime d'abord de l'enthousiasme, mais devient insatisfaite en découvrant une différence de prix significative entre l'achat sur Amazon et l'entrepôt alimentaire local. La déclaration "Wow, Amazon se fait vraiment de l'argent avec celui-ci" et la recommandation de "Magasiner localement avant d'acheter cette merveille hors de prix" indiquent un sentiment négatif.

Figure 13: Limites de RoBERTa : Example 1

```
l_df.query("Score == 5").sort_values("rta_neg" , ascending=False)['Text'].values[1]
```

```
"I've eaten other brands of unsalted potato chips and they've always been terrible.  I didn't expect much when I bought thes
e.  But after I tasted them I was amazed.  I have never enjoyed unsalted potato chips before these.  They have a natural pot
ato flavor with thick, crunchy chips.  The only bad thing is the bag is a pain to open.  Use scissors and save yourself the
aggravation."
```

Le sentiment dans ce texte est positif. La personne avait initialement des attentes basses en se basant sur des expériences passées avec d'autres marques de chips de pommes de terre non salées, mais a été agréablement surprise et émerveillée après avoir essayé celles-ci. Les aspects positifs mentionnés comprennent une saveur naturelle de pomme de terre avec des chips épaisses et croquantes. Cependant, un problème mineur est noté avec la difficulté d'ouvrir le sac, et la suggestion est donnée d'utiliser des ciseaux pour éviter l'aggravation. Dans l'ensemble, le ton est favorable.

Figure 14: Limites de RoBERTa : Example 2

Suggestions et améliorations possibles du projet

- Exploration de sous-catégories alimentaires : Explorer les sentiments dans des sous-catégories spécifiques de produits alimentaires pour obtenir des informations plus détaillées.
- Enrichissement du modèle avec des données spécifiques au domaine : Utiliser des données spécifiques au domaine alimentaire pour enrichir le modèle.
- Intégration de la rétroaction des entreprises : Permettre aux entreprises de répondre aux avis pourrait offrir une perspective plus complète sur la satisfaction du client.
- Analyse de la dynamique temporelle : Investiguer les tendances au fil du temps en analysant les changements mensuels dans les avis.
- Comparaison avec d'autres modèles NLP : Comparer avec d'autres modèles NLP pour évaluer la performance relative.

Implications et conclusions tirées des résultats obtenus

- Tendance positive globale : La distribution des scores suggère une tendance positive globale dans les avis des clients sur les produits alimentaires d'Amazon.
- Importance des avis avec des scores élevés : La majorité des scores sont entre 4 et 5, soulignant l'importance des avis positifs.
- Prévalence des évaluations positives : La fréquence élevée des évaluations positives, en particulier avec un score de 5, peut indiquer une propension des utilisateurs à partager leurs expériences positives.

Récapitulation des Principaux Résultats

- Tendance Globale Positive : L'analyse des scores des avis suggère une tendance globale positive, avec une prédominance d'évaluations élevées, principalement de 4 et 5.
- Identification de Tendances Alimentaires : RoBERTa a été utilisé avec succès pour identifier des tendances émergentes dans les avis, notamment des préférences alimentaires, des aspects appréciés ou critiqués spécifiques, et des évolutions au fil du temps.
- Importance des Avis Positifs : Les avis positifs avec des scores élevés sont prédominants, indiquant que la satisfaction des clients est généralement élevée.

Réponse à la Question Initiale

La question initiale visait à comprendre les avis des clients sur les produits alimentaires d'Amazon. Les résultats obtenus suggèrent que la majorité des clients expriment des sentiments positifs envers ces produits. La satisfaction semble être élevée, ce qui peut être une information précieuse pour les entreprises cherchant à améliorer leurs produits.

Contributions du Projet à la Connaissance du Domaine

- Analyse Fine des Sentiments : L'utilisation de RoBERTa a permis une analyse fine des sentiments exprimés dans les avis, offrant une compréhension approfondie des opinions des clients.
- Identification de Tendances et de Préférences : Le projet a contribué à l'identification de tendances émergentes et de préférences alimentaires, offrant ainsi des informations utiles pour les entreprises cherchant à répondre aux attentes du marché.
- Utilisation de Données Réelles d'Amazon : En utilisant un ensemble de données provenant des avis réels des clients sur Amazon, le projet a contribué à une analyse basée sur des données concrètes, renforçant ainsi la validité des résultats.

Conclusion Finale

En conclusion, ce projet a fourni des informations précieuses sur les sentiments des clients à l'égard des produits alimentaires d'Amazon, mettant en lumière des tendances, des préférences et des points forts. Ces connaissances peuvent être exploitées par les entreprises pour améliorer leurs produits et satisfaire davantage leurs clients.

Références

- Xinyue Zhao et Yuandong Sun. (2022). ”Amazon Fine Food Reviews with BERT Model.” Procedia Computer Science, Volume 208, Pages 401-406. DOI: 10.1016/j.procs.2022.10.056. Elsevier B.V. Available online at: https://www.sciencedirect.com/science/article/pii/S1877050922030215.
- Hugging Face. (2022). Documentation Transformers - Modèle RoBERTa. https://huggingface.co/docs/transformers/model_doc/roberta.
- Mulla, R. (2022). Projet d’analyse de sentiment en Python avec NLTK et Transformers. Classifiez les critiques d’Amazon !! [Vidéo]. YouTube. https://www.youtube.com/watch?v=QpzMWQvxXWk.
- Robikscube. (2022). Analyse de sentiment en Python [Didacticiel sur YouTube]. Kaggle. https://www.kaggle.com/code/robikscube/sentiment-analysis-python-youtube-tutorial/notebook.