Amazon Review Analysis

Phase 1

Ismaël SYLLA

17 décembre 2025

Plan

- Introduction
- Identification du besoin
- Veille technologique et réglementaire
- Sélection des données
- Prototype
- La solution
- Conclusion

Introduction

Introduction

amazon

Acteur influent dans le e-commerce et le cloud computing.

- ✓ 3ème mondiale (Interband).
- ✓ Réseau logisitique performant
- ✓ Présence dominante dans le e-commerce et cloud computing.
- ✓ Large base de clientèle fidèle.

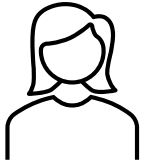
- ✓ Dépendance aux vendeurs tiers.
- ✓ Problèmes juridiques.
- ✓ Délai de livraison important dans certaines zones rurales.
- ✓ Vulnérabilité aux cyberattaques.

- ✓ Croissance continue dans les marchés émergeants.
- ✓ **Valorisation des retours clients.**
- ✓ Développement durable et innovation logistique.

- ✓ Marché très concurrentiel.
- ✓ Pression réglementaire.
- ✓ Multiplication des cyberattaques.

Identification du besoin

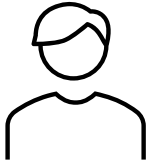
Besoin identifié



- Léa
- 31 ans
- CDI et maman

Besoin identifié : Avoir les avis les pertinents dans la file des avis pour chaque produit.

Cas d'utilisation : Mettre en avant les avis produits les plus pertinents.



- Mathis
- 42 ans
- entrepreneur

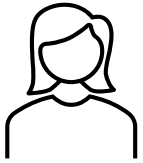
Besoin identifié :

- Améliorer la visibilité et la performance de ses produits.
- Comprendre les critères influençant le classement des produits.

Cas d'utilisation :

- Analyse des avis clients et tendances de satisfaction.
- Transparence sur les algorithmes de classement produits.

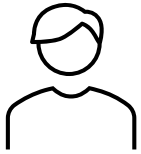
Besoin identifié



- Emma
- 21 ans
- Etudiante en psychologie

Besoin identifié : Bénéficier d'un service de livraison rapide.

Cas d'utilisation : Optimisation logistique et amélioration de la livraison dans les zones rurales.
.



- Rayan
- 25 ans
- Business Analyst chez Amazon

Besoin identifié :
- Pouvoir automatiser l'analyse et la classification des avis clients sur les produits.

Cas d'utilisation :
- Classer les avis et identifier leur pertinence.

Besoin identifié

Besoin retenu :
Classer les avis et identifier leur pertinence.

Valeurs ajoutées :
- Eliminer un travail manuel et gagner du temps.
- Avoir une analyse plus rapide et plus fiable.
- Ce cas d'usage peut être considéré comme un levier et une première brique pour d'autres besoins.

Veille technologique et réglementaire

Veille technologique et réglementaire

Amazon S3 : devient une vraie brique de data lake intelligente avec des optimisations automatiques des coûts. Idéal pour le stockage massif, historisation et gouvernance de données brutes.
Snowflake : datawarehouse plus rapide et plus sécurisés, il renforce la data gouvernance et la perfomance analytique sur une seule plateforme.
Databricks : plateforme lakehouse (combine data lake + data warehouse + AI/ML) donne une meilleure production des modèles, idéal pour industrialiser le ML et data quality à grande échelle.
Amazon SageMaker : plateforme de ML managée (entrainement modèles, dépoilement de modèles). Permet un meilleur contrôle de données.
Zero-shot : classification de texte sans données d'entraînement spécifiques, très rapide à déployer.
VADER ; analyse de sentiment efficace pour les avis clients.

Veille technologique et réglementaire

Veille réglementaire :

- RGPD : demeure la référence sur consentement, minimisation, anonymisation ou pseudonymisation des données à caractères personnelles.
- CCPA : L'équivalent californien du RGPD : impose le droit à l'information, à la suppression et à l'opposition à la vente des données personnelles.
- AI Act (Régment UE) : L'AI Act impose des obligations de transparence (déclarations si contenu IA), restrictions sur la manipulation.
- RSE : encourage une IA éthique et sobre en évaluant l'empreinte carbone des modèles, choix cloud responsables, et inclusion sociale dans les projets IA. Tendance forte en 2025.

Veille technologique et réglementaire

Le partage de la veille technologique et réglementaire :

- ✓ Partage complet dans de la documentation Confluence.
- ✓ Partage via un canal dédié sur Teams.

Sélection des données

Sélection des données

- REVIEW : Récupérer tous les avis des clients.
- PRODUCT_REVIEW : Récupérer les liens entre produit et avis.
- PRODUCT : récupérer les produits.
- CATEGORY : récupérer les catégories des produits.
- REVIEW_IMAGES : récupérer les avis qui ont des images.
- BUYER : identifier l'acheteur.

Pourquoi cette phase ?

- Mieux comprendre la structure des données.
- Identifier les tables nécessaires au cas d'utilisation identifié.
- Commencer à visualiser la solution.

* SHIPMENT : table très intéressante avec des informations riches. Mais sa volumétrie est de 5 lignes.

Prototype

Prototype

Phase 1 : Comprendre les données

1 produit avec 376 avis

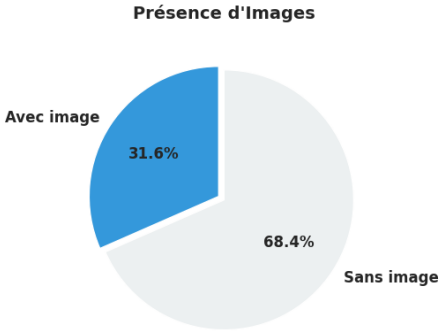
- 73,1 % des avis ont 5 étoiles.
- 28,5% des avis ont une photo.
- Médiane sur la distribution du texte : 58 caractères.
- Moyenne sur la distribution du texte : 160 caractères.

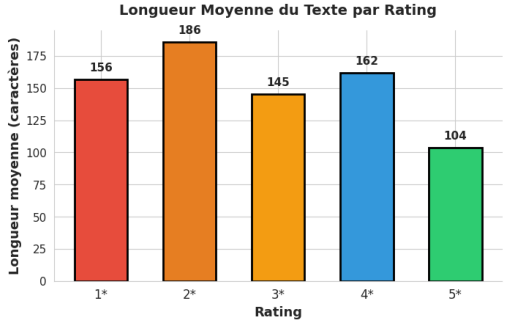
10 produits avec 2592 avis

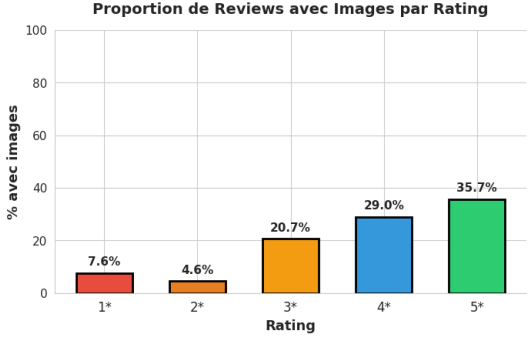
- 76,4 % des avis ont 5 étoiles.
- 31,6 % des avis ont une photo.
- Médiane sur la distribution du texte : 44 caractères.
- Moyenne sur la distribution du texte : 117 caractères.

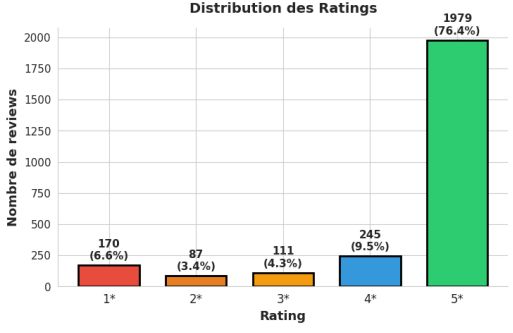
Prototype

Phase 1 : analyse sur 10 produits (2592 avis)









RÉSUMÉ STATISTIQUE

- Reviews totales : 2592
- Avec images : 818 (31.6%)
- Avec commandes : 2592 (100.0%)
- Longueur texte :
  - Moyenne : 117 caractères
  - Médiane : 44 caractères
  - Min : 4 | Max : 1815
- Rating moyen : 4.46/5

Prototype

Phase 2 : Appliquer un algorithme de pondération

Calculer la pertinence :

- 30 % sur la longueur du texte.
- 20 % sur la présence d'une image dans l'avis.
- 10 % si celui qui a commenté a effectué un achat.
- 15 % sur la note laissée si elle est de 1 ou 5.
- 25 % sur les mots utilisés dans le commentaire.

La solution

La solution

Spécifications fonctionnelles et techniques

Objectifs :
Développer une solution automatisée pour classer les avis et détecter les avis les plus pertinents.

Contraintes :
- Jeu de données limité.
- Infrastructure favoriser les solutions gratuites/open source ou les moins coûteuses.

Considérations réglementaires et éthiques :
- Respect des réglementations (RGPD, CCPA).
- Anonymisation des données à caractères personnelles.
- Neutralité des critères (dans la pondération).

Conclusion

Conclusion

Présentation des résultats des prototypes :
Les résultats avec Stats
Les limites (les étoiles positifs…)
Perspectives pour y remédier introduire une approche plus robuste, avec les technos identifiés dans la veille techno + zéro shot
Après passage à la phase 2

Merci !