

Analyser et catégoriser les avis



Projet : Amazon Review Analysis

Auteur : Dyhia TOUAHRI

Date : 12 novembre 2025

Version 1.0

Table des matières

1.	Introduction	3
1.1.	Objectif du projet	3
1.2.	Résumé du pipeline ETL	3
2.	Méthodologie	4
2.1.	Approche Zero-Shot Classification	4
2.2.	Algorithme de pondération	4
3.	Définition des seuils.....	6
3.1.	Observation des distributions	6
3.2.	Tests multi-seuils (TOP 25%, 40%, 45%, 50%)	6
3.3.	Seuils finaux retenus.....	6
4.	Analyse des résultats	7
4.1.	Distribution du statut de pertinence.....	7
4.2.	Scores moyens par statut	7
4.3.	Pertinence par catégorie	7
4.4.	Pertinence par produit	7
4.5.	Longueur du texte.....	8
4.6.	Impact du rating extrême	8
4.7.	Influence des mots-clés	8
4.8.	Pertinence selon le rating.....	8
4.9.	Impact de la présence d'image	9
4.10.	Impact de l'achat vérifié (HAS_ORDERS)	9
4.11.	Longueur des descriptions des avis pertinents	9
5.	Conclusion.....	10

1. Introduction

1.1. Objectif du projet

L'objectif de ce projet est de construire une solution capable d'identifier automatiquement si un avis client Amazon est pertinent ou non pertinent, en combinant :

- Des techniques d'analyse NLP (Zero-Shot Classification),
- Un algorithme de pondération basé sur plusieurs signaux textuels,
- Un pipeline ETL complet permettant d'ingérer, transformer et analyser les données.

L'enjeu final est de fournir un ensemble d'avis fiables et de haute qualité permettant :

- Des analyses de satisfaction plus robustes,
- Des segmentations par produit ou par catégorie,
- Des insights exploitables pour améliorer l'expérience client.

1.2. Résumé du pipeline ETL

Le pipeline développé comporte les étapes suivantes :

1. Extraction
 - Récupération des avis clients (texte, note, métadonnées).
2. Transformation
 - Nettoyage et normalisation.
 - Calcul de nouvelles features :
 - Longueur du texte,
 - Présence image,
 - Achat vérifié,
 - Score Zero-Shot (catégorisation),
 - Score de confiance du modèle,
 - Score de mots-clés,
 - Score de pertinence global.
3. Chargement
 - Stockage de la table transformée et enrichie.
 - Création d'une table d'analyse REVIEW_RELEVANT.
4. Analyse
 - Application d'un modèle décisionnel basé sur des seuils.
 - Construction d'un dashboard Streamlit dans Snowflake.

L'ensemble de la chaîne permet une automatisation complète de l'identification des avis pertinents.

2. Méthodologie

2.1. Approche Zero-Shot Classification

Pour analyser automatiquement le contenu des avis clients, nous avons utilisé une approche de Zero-Shot Classification, basée sur un modèle NLP avancé :

« mDeBERTa-v3-base-xnli-multilingual », hébergé sur Hugging Face.

Ce modèle est l'un des plus performants pour la classification NLI multilingue et permet d'interpréter du texte dans plus de 100 langues, dont le français, l'anglais, l'espagnol, etc.

Pourquoi Zero-shot ?

Aucun entraînement spécifique n'est nécessaire.

On fournit simplement des catégories, et le modèle détermine laquelle correspond le mieux à l'avis.

Cela évite :

- la création d'un dataset annoté,
- la phase d'entraînement longue,
- les risques de surapprentissage.

Dans ce projet, nous avons défini les 4 catégories représentant les principaux motifs d'un avis Amazon :

- *product quality or satisfaction*
- *product defect or damaged item*
- *customer service issue*
- *shipping or packaging problem*

Le modèle fournit :

- une catégorie prédite,
- un score de confiance (CONFIDENCE_SCORE).

Ce score mesure la certitude du modèle quant à l'interprétation du texte.

2.2. Algorithme de pondération

En complément du Zero-Shot, nous avons conçu un algorithme de scoring pondéré permettant d'estimer la pertinence intrinsèque d'un avis.

Ce score, appelé **RELEVANCE_SCORE**, repose sur plusieurs composants mesurant la qualité du texte ou la fiabilité de l'auteur.

Le but est d'obtenir une évaluation plus humaine et plus fine qu'une simple classification NLP.

Nous calculons un score entre **0 et 100** basé sur 5 dimensions :

Composant	Description	Poids
TEXT_LENGTH_SCORE	Qualité du texte via sa longueur	30,00%
HAS_IMAGE	Présence d'une preuve visuelle	20,00%
HAS_ORDERS	Achat vérifié	10,00%
IS_EXTREME_RATING	Avis très positif ou négatif	15,00%
KEYWORD_SCORE	Sentiment et qualité linguistique	25,00%

- TEXT_LENGTH_SCORE : nous utilisons une fonction gaussienne centrée sur 300 caractères.
- IS_EXTREME_RATING : les avis extrêmes sont souvent plus détaillés, plus argumentés, plus émotionnels et plus informatifs.
- KEYWORD_SCORE : Au lieu d'un simple score de mots-clés, nous utilisons VADER, un outil NLP spécialisé dans le sentiment pour les textes courts.
- HAS_IMAGE : un avis avec image est plus engageant, plus vérifiable et souvent plus pertinent sur un produit physique.
- HAS_ORDERS : les avis associés à un achat réellement effectué sont souvent plus fiables, moins suspects et plus objectifs.

La formule finale :

```
# Compute the final relevance_score using weighted components
df_reviews['RELEVANCE_SCORE'] = (
    0.30 * df_reviews['TEXT_LENGTH_SCORE'] +
    0.20 * df_reviews['HAS_IMAGE'] +
    0.10 * df_reviews['HAS_ORDERS'] +
    0.15 * df_reviews['IS_EXTREME_RATING'] +
    0.25 * df_reviews['KEYWORD_SCORE']
) * 100 # scale to 0-100
```

Les deux systèmes (Zero-Shot + pondération) sont ensuite combinés pour décider si un avis est classé comme pertinent ou non pertinent.

3. Définition des seuils

L'analyse de distribution des scores et l'étude des pourcentages d'avis conservés selon différents seuils ont permis de choisir les valeurs optimales.

3.1. Observation des distributions

Confidence Score

- Pic principal ≈ 95 .
- Distribution majoritairement entre 70 et 100.
- Le modèle Zero-Shot est globalement très sûr.

Relevance Score

- Pic autour de 60.
- Forte dispersion → variabilité élevée de pertinence textuelle.
- Creux autour de 80-85 → zone potentielle pour un seuil strict.

Densité faible (≈ 0.035)

→ Les avis sont très hétérogènes.

3.2. Tests multi-seuils (TOP 25%, 40%, 45%, 50%)

Le principe de ce test est d'analyser combien d'avis seraient labellisés « PERTINENT » si on prenait le top 25%, 40%, 45% ou 50% de score de pertinence ou de confiance et ainsi voir combien d'avis seraient exclus.

Exemple :

Test	Seuil CONFIDENCE	Seuil RELEVANCE	Avis conservés
Top 25%	89.97	68.07	8.7%
Top 50%	74.95	58.75	30.3%
Top 45% & Top 50%	78.3	58.8	27.8%

Principales observations :

1. Confidence Score est systématiquement plus élevé que Relevance Score.
2. Le passage de 25% → 50% fait passer de 8.7% à 30.3%, énorme écart.
3. Les seuils intermédiaires (40–50%) offrent un bon compromis.
4. Le couple 78.3% / 58.8% conserve 27.8% des données → bonne balance entre qualité et volume.

3.3. Seuils finaux retenus

Les seuils recommandés et utilisés pour la classification sont :

- **CONFIDENCE_SCORE $\geq 78.3\%$**
- **RELEVANCE_SCORE $\geq 58.8\%$**

Ces seuils retiennent ~31 000 avis (**27.8%**) considérés comme fiables et pertinents.

4. Analyse des résultats

Les analyses suivantes proviennent directement du dashboard analytique.

4.1. Distribution du statut de pertinence

Moins d'un tiers des avis sont pertinents selon les seuils retenus.

Statut	Nombre	Pourcentage
RELEVANT	30896	27.75%
IRRELEVANT	80426	72.25%

4.2. Scores moyens par statut

Les avis pertinents sont plus longs, mieux scorés et surtout beaucoup plus confiants.

Statut	Relevance	Confidence	Text Length Score
IRRELEVANT	55.05	65.60	0.55
RELEVANT	70.26	90.81	0.71

4.3. Pertinence par catégorie

Les deux catégories sont proches, mais Premium Beauty produit légèrement plus d'avis pertinents.

Catégorie	Nb avis	% pertinents
All Beauty	110018	0,28
Premium Beauty	1304	0,31

4.4. Pertinence par produit

Les extraits montrent une grande variabilité :

- Certains produits ont **0%** d'avis pertinents.
- Certains montent à **40%**.
- La plupart se situent entre **10% et 30%**.

Les produits génèrent des avis très inégaux en qualité → potentiel d'analyse produit par produit.

Analyse 4 — Pertinence par produit

Objectif : Identifier les produits à très fort / faible taux de pertinence.

P_ID	PRODUCT_NAME	NB_REVIEWS	NB_RELEVANT	PCT_RELEVANT	AVG_RATING
0	B0BM4GX6TT	376	82	21.81	4.4043
1	B085B7B1M	350	91	26	4.6629
2	B00XK20Y4C	341	19	5.57	4.7742
3	B012Q9NGE4	238	34	14.29	4.3824
4	B08D8SD29C	238	0	0	4.6975
5	B07C533YCW	215	86	40	4.4837
6	B07T2L6JQR	215	8	3.72	4.8558
7	B019GBG0IE	209	58	27.75	3.5646
8	B09X9BG4FC	207	79	38.16	4.4396
9	B08LSKN7X4	203	51	25.12	3.9557

4.5. Longueur du texte

Les avis pertinents sont légèrement plus longs, mais la différence reste modérée → cela confirme que la longueur n'est pas un critère unique, mais un bon indicateur complémentaire.

Statut	Moyenne	Min	Max
RELEVANT	188	5	10221
IRRELEVANT	170	1	14643

4.6. Impact du rating extrême

Les avis extrêmes (1★ ou 5★) contiennent beaucoup plus d'avis pertinents. C'est logique : les avis les plus émotionnels sont souvent plus détaillés.

Rating extrême	Pertinent	Non pertinent
false	2186	26315
true	28710	54111

4.7. Influence des mots-clés

Les mots clés sont un excellent indicateur de pertinence.

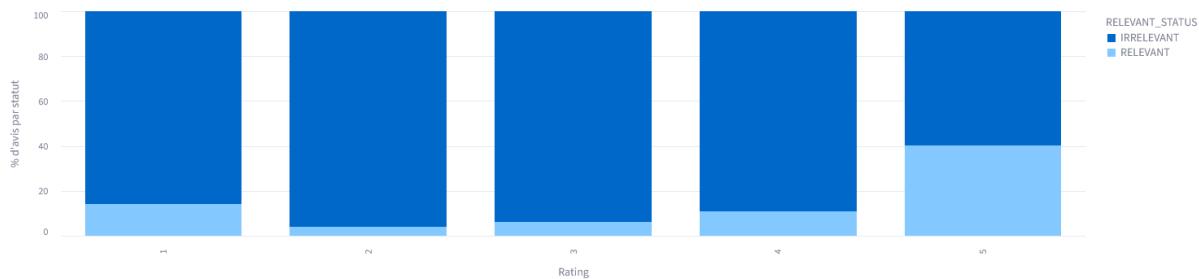
Statut	Moyenne Keyword Score
RELEVANT	0.87
IRRELEVANT	0.66

4.8. Pertinence selon le rating

Les avis pertinents augmentent fortement avec le nombre d'étoiles :

- À 1★ → seulement 14% pertinents
- À 5★ → 40% pertinents

Les avis 5★ sont plus riches et descriptifs que les avis négatifs courts.



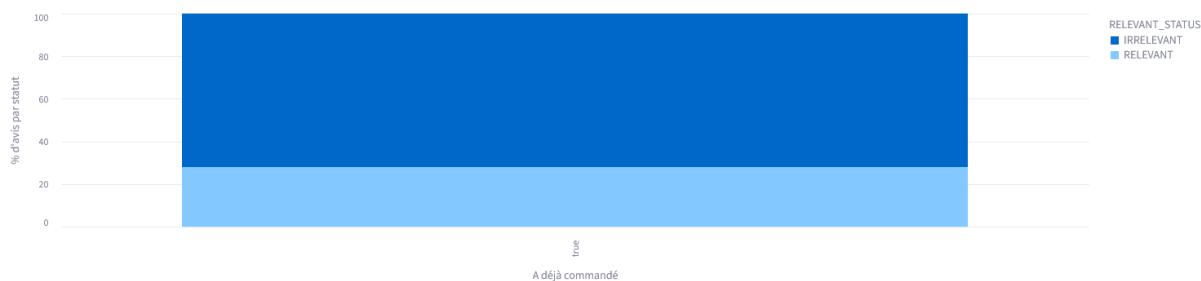
4.9. Impact de la présence d'image

Les avis avec image sont nettement plus pertinents.

Image	% pertinents
Sans	0,26
Avec	0,41

4.10. Impact de l'achat vérifié (HAS_ORDERS)

On constate que tous les achats sont vérifiés. Cela représente donc 28% des avis. Tous les avis labélisés pertinents ont le has_orders = vrai.



4.11. Longueur des descriptions des avis pertinents

La grande majorité des avis pertinents font **plus de 100 caractères**.

Le modèle identifie donc que les avis trop courts ne sont pas pertinents.

Niveau	Description	Compte
LEVEL0	0–5 char	3
LEVEL1	6–10	36
LEVEL2	11–20	186
LEVEL3	21–40	1235
LEVEL4	41–100	7898
LEVEL5	> 100	21538

5. Conclusion

Ce travail a permis de construire une solution complète pour détecter automatiquement des avis pertinents grâce à :

- un pipeline ETL fiable,
- une combinaison NLP + pondération,
- une méthodologie rigoureuse de choix des seuils,
- un dashboard d'analyse interactif dans Snowflake.

Les seuils retenus (**Confidence $\geq 78.3\%$** et **Relevance $\geq 58.8\%$**) permettent de conserver un ensemble d'environ **31 000 avis de haute qualité**, adaptés pour des analyses fiables.

Les analyses montrent que :

- les avis pertinents sont plus longs, plus confiants et davantage associés à des mots clés,
- les avis extrêmes (positifs ou négatifs) sont souvent plus détaillés,
- la présence d'images augmente fortement la pertinence,
- certaines catégories et produits produisent des avis plus qualitatifs que d'autres.

Cette méthodologie peut être réutilisée et enrichie pour :

- suivre l'évolution de la qualité des avis dans le temps,
- détecter automatiquement les produits posant problème,
- améliorer l'expérience client via des insights concrets.