

Analyser et catégoriser les avis



Projet : Amazon Review Analysis

Auteur : Ismaël SYLLA

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.....	11

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.
- Adaptation de l'application Streamlit des Business Analysts.

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 se focalisent sur les features utilisées pour les algorithmes.

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	Godefroy Tint Kit for Spot Coloring, Dark Brown	376	82	21.81	4.4043
1	B085B7B1M	Salux Nylon Japanese Beauty Skin Bath Wash Cloth/towel (3) Blue Yellow and Pink	350	91	26	4.6629
2	B00XK20Y4C	Soft 'N Style Butterfly Clamps, Assorted Colors, 1 Dozen	341	19	5.57	4.7742
3	B012Q9NGE4	Moroccan Argan Oil Shampoo, 8 Fl Oz - Smooths and Repairs - Sulfate Free - Natural - Imported from Morocco	238	34	14.29	4.3824
4	B08D8SD29C	Gortin Boho Headbands Stretch Wide Head Bands Blcak Butterfy Hair Bands Yoga Turban Headband	238	0	0	4.6975
5	B07C533YCW	Segbeauty empty bottle 160083	215	86	40	4.4837
6	B07T2L6JQR	Tea Tree Lemon Sage Thickening Liter Duo Set	215	8	3.72	4.8558
7	B019GBG0IE	Collapsible Hair Diffuser by The Curly Co. with The Curly Co. Satisfaction Guarantee	209	58	27.75	3.5646
8	B09X9BG4FC	Makone Crystal Crowns and Tiaras with Comb Headband for Girl or Women Birthday Party Wedding	207	79	38.16	4.4396
9	B08LSKN7X4	Meeteasy Dental Cleaner Tool Kit - Dental Care for Adult - 100% Proven Safe	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.

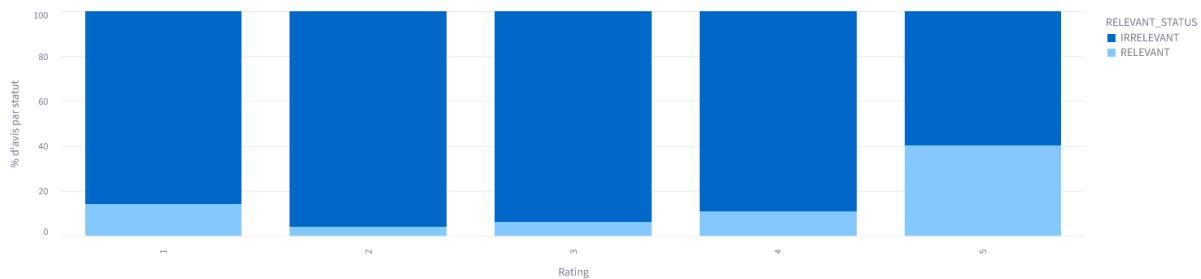
Statut	Moyenne Keyword Score
RELEVANT	0.87
IRRELEVANT	0.66

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. Indicateurs clés de performance (KPI)

Amazon Review Analysis

KPI 1 — Top 20 produits avec le plus d'avis pertinents

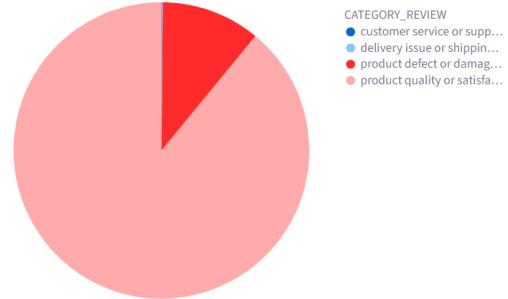
Produits générant le plus d'avis classés RELEVANT.

	PRODUCT_NAME	NB_RELEVANT_REVIEWS
0	Salux Nylon Japanese Beauty Skin Bath Wash Cloth/towel (3) Blue Yellow and Pink	91
1	Segbeauty empty bottle 160083	86
2	Godefroy Tint Kit for Spot Coloring, Dark Brown	82
3	Makone Crystal Crowns and Tiaras with Comb Headband for Girl or Women Birthday Party Wedding Prom Bridal Christmas Valentine... (03 Pink)	79
4	Picoway 20 Pack Mouse Ears Solid Black and Red Bow Headband	75
5	4 Point Eyebrow Pen, Micro Ink Tat Brow Pen Waterproof Eyebrow Pencil With Micro-Fork Tips for Daily Natural Eye Brown Makeup (Dark Brown/ Chestnut)	64
6	Bed Head Curve Check Curling Wand for Tousled Waves and Texture, Jumbo Barrel	59
7	Collapsible Hair Diffuser by The Curly Co. with The Curly Co. Satisfaction Guarantee	58
8	Foot Peel Mask 3 Pack Exfoliating Foot Mask for Callus and Dry Dead Skin Repair Rough Heels Get Baby Soft Smooth Touch Feet for Man Woman (Lavender Rose Aloe)	55
9	Poly Gel Nail Kit, obove Nail Builder Gel Extension Nail thickening Solution French Nail Art Equipment - 6 Colors 15ML	52

KPI 2 — Répartition par catégorie review

Distribution des avis pertinents selon la catégorie review.

	CATEGORY REVIEW	NB_RELEVANT
0	product quality or satisfaction	27510
1	product defect or damaged item	3344
2	delivery issue or shipping delay	22
3	customer service or support	20



KPI 3 — Top 20 clients (avis pertinents)

Analyse des clients avec le plus d'avis pertinents.

	BUYER_ID	NB_RELEVANT_REVIEWS
0	85efb6899650317220f160342a96209abf1cf926097783d3367151da173118c3	32
1	f74834ce78f2f03fde0d9e465c267c6dc19eae5b160c154a0ab4d52d094ffdd	30
2	e64b9219a55d43883559c49f580b14497f702406630c3b608e7ae1314823c835	29
3	578a4900bbf4fbca23eef4389a62981ed34492863b6ce5e8737812bb22a6edf8	20
4	275f78c683e565c5f02b50bfd8bbd1e61918adff7d37a0f7688bec5a0403b541e	20
5	add851ebcb45ab126a7165be3aac7f63536942d0508d0e0a8f110211d1fe4d	13
6	fd446b2ea78c2efdf50006993c3d8f3679af71be166e31e43fc60fff9b650	13
7	1557bc15d31cf4ea4abb6a2b105c9302cf06d9b881299d0b5d63a967f63c4095	12
8	06dcfb93ece90e3759d1a222a87d2d1ab7751709cf9611529d2a1070d457e280	12
9	b064719d795e73d7b1e903a8fb75fc8f32ed102643358f8b14595c3982c9aa4e	10

Amazon Review Analysis

KPI 4 — Les 20 produits avec le moins d'avis pertinents

Produits avec le moins de reviews.

	PRODUCT_NAME	NB_REVIEWS
0	LILYCUTE gel nail polish kit (set1)	1
1	FANXITON 8 Styles Mixed False Eyelashes Fluffy Natural False Lashes 3D Volume Faux Mink Lashes Pack 16 Pairs	1
2	Bloom Clarity Hydrating Body Mist	1
3	SHILLS Black Mask, Peel Off Mask, Blackhead Remover Mask, Charcoal Mask, Blackhead Peel Off Mask, Plus Activated Charcoal Konjac Sponge	1
4	Laerdal Speedblocks Head Immobilizer Strap/pad - Model 983096 - Set	1
5	Silver Plated Bead Chain Anklet Bracelet Foot Jewelry Barefoot Sandal Beach Woman	1
6	Ombre Jumbo Braiding Hair Extensions Synthetic Yaki Straight High Temperature Fiber Crochet Braids Hairstyles (Blond Pink Purple)	1
7	Dove Men Plus Care Everyday Gift Pack, Clean Comfort	1
8	QtGirl Mermaid Sequin Headbands for Girls, 3 Pack Reversible Flip Sequins Headband Stretch Elastic Hairband for Teens Girls and Women Party	1
9	First Aid Beauty Ultra Repair Sampler Set: Vanilla Citron & Original (2oz Jars)	1

KPI 5 — Top 20 clients les plus actifs

Nombre total d'avis + ratio de pertinence.

	BUYER_ID	TOTAL_REVIEWS	NB_RELEVANT	PCT_RELEVANT
0	578a4900bbf4fbca23eeff4389a62981ed34492863b6ce5e8737812b22a6edf8	115	20	0.174
1	85efb6899650317220f160342a96209abf1cf926097783d3367151da173118c3	79	32	0.405
2	275f78c683e565c5f02b50bfd8bbd1e61918adf7d37a0f7688bec5a0403b541e	65	20	0.308
3	e64b9219a55d43883559c49f580b14497f702406630c3b608e7ae1314823c835	62	29	0.468
4	1557bc15d31cfb4ea4bb6a2b105c9302cf06d9b8b1299d0b5d63a967f63c4095	47	12	0.255
5	f74834ce78f2f03fde0d9e465c267c6fdc19eae5b160c154a0ab4d52d094fdd9	40	30	0.75
6	1a874615a6b50357b573d0a044ffb30a9d0f899df7d303df6a169d890c96564	34	5	0.147
7	9f7cb27hb28308169f7d4e309f985602b47dc837f32dec2a2056b9d1b96c0603	30	8	0.267
8	8865838784a8ac3a351d4820203b15081312b28ca06fefa752388b8eb8088099	29	2	0.069
9	68c9c139ea768b7047c4c2fe06a1ebfe8bbe07e66562ee52e94ab97e1ef8c28f	29	10	0.345

KPI 6 — Produits les plus commentés

Produits générant le plus d'engagement client.

	PRODUCT_NAME	NB_REVIEWS
0	Godefroy Tint Kit for Spot Coloring, Dark Brown	376
1	Salux Nylon Japanese Beauty Skin Bath Wash Cloth/towel (3) Blue Yellow and Pink	350
2	Soft 'N Style Butterfly Clamps, Assorted Colors, 1 Dozen	341
3	Tea Tree Lemon Sage Thickening Liter Duo Set	280
4	Moroccan Argan Oil Shampoo, 8 Fl Oz - Smooths and Repairs - Sulfate Free - Natural - Imported from Morocco by Pure Body Naturals	238
5	Gortin Boho Headbands Stretch Wide Head Bands Blcak Butterfly Hair Bands Yoga Turban Head Wraps Fashion Head Scarfs for Women and Girl	238
6	Segbeauty empty bottle 160083	215
7	Collapsible Hair Diffuser by The Curly Co. with The Curly Co. Satisfaction Guarantee	209
8	Makone Crystal Crowns and Tiaras with Comb Headband for Girl or Women Birthday Party Wedding Prom Bridal Christmas Valentine... (03 Pink)	207
9	Meeteasy Dental Cleaner Tool Kit - Dental Care for Adult - 100% Proven Safe	203

6. 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.