# PROJET DE CLASSIFICATION DE COMMENTAIRES

## COMMENTAIRES AMAZON

---

**Auteurs :**

- Salimou ABDOULAYE HALIDOU
- Parfait Jemmy Prodige NGOY
- Ibrahima Fa LO

**Date :** Juin 2025
**Cours :** Webscrapping

---

## RÉSUMÉ EXÉCUTIF

Ce projet développe un système complet d'analyse de commentaires pour les commentaires clients d'Amazon. Il comprend la collecte automatisée de données, le préprocessing spécialisé pour le français, l'entraînement de modèles de machine learning, et le déploiement via une application web interactive.

---

## TABLE DES MATIÈRES

1. Collecte et Acquisition des Données
2. Analyse Exploratoire et Préparation
3. Développement du Preprocessing
4. Vectorisation TF-IDF
5. Évaluation Comparative des Modèles
6. Création du Pipeline Final
7. Développement Application Web
8. Optimisation et Déploiement
9. Résultats et Performances
10. Technologies Utilisées
11. Innovation et Contributions
12. Conclusion

---

## 1. COLLECTE ET ACQUISITION DES DONNÉES

### 1.1 Outil Développé

**Spider Scrapy personnalisé** : amazon_spider_bv.py

## 1.2 Stratégie de Collecte

**Approche multi-catégories** couvrant 10 domaines principaux d'Amazon :

- Électronique & High-Tech
- Maison & Jardin
- Mode & Beauté
- Sport & Loisirs
- Jouets & Jeux
- Auto & Moto
- Livres & Culture
- Santé & Bien-être
- Animaux
- Alimentation & Boissons

## 1.3 Caractéristiques Techniques

- **Scraping éthique** : Délais respectueux (4s), rotation d'user-agents
- **Échantillonnage aléatoire** : Diversité maximale des produits et commentaires
- **Classification automatique** : Commentaires et catégorie de produit
- **Gestion robuste** : Retry automatique, gestion des erreurs

## 1.4 Volume de Données Collectées

- **Objectif** : 1000 produits uniques
- **Résultat** : 5021 commentaires collectés
- **Format** : CSV + JSON avec métadonnées enrichies

## 1.5 Structure des Données

| Colonne | Description |
|---|---|
| asin | Identifiant unique du produit |
| product_title | Nom du produit |
| review_rating | Note (1-5 étoiles) |
| review_text | Texte du commentaire |
| review_sentiment | Classification automatique des commentaires |
| verified_purchase | Achat vérifié |
| helpful_votes | Votes d'utilité |
| date | Date du commentaire |

## 2. ANALYSE EXPLORATOIRE ET PRÉPARATION

### 2.1 Distribution Initiale

- **Commentaires positifs** : 90.6% (notes 4-5) - 4548 commentaires
- **Commentaires neutres** : 4.7% (note 3) - 237 commentaires
- **Commentaires négatifs** : 4.7% (notes 1-2) - 236 commentaires

### 2.2 Problème Identifié

Déséquilibre extrême des classes nécessitant une stratégie d'équilibrage.

### 2.3 Stratégie d'Équilibrage

- **Sous-échantillonnage** des commentaires positifs (50% conservés)
- **Conservation totale** des commentaires neutres et négatifs

### 2.4 Résultat Final

**Dataset équilibré** : 2747 commentaires

- Positifs : 2274 (82.8%)
- Neutres : 237 (8.6%)
- Négatifs : 236 (8.6%)

---

## 3. DÉVELOPPEMENT DU PREPROCESSING

### 3.1 Approche Spécialisée Français

Système de préprocessing adapté spécifiquement au français d'Amazon.

### 3.2 Fonctionnalités Principales

- Nettoyage des URLs, emails et caractères spéciaux
- Suppression des stopwords français et termes spécifiques Amazon
- Stemming adapté à la langue française
- Tokenisation optimisée pour les commentaires clients
- Normalisation des accents et de la casse

---

## 4. VECTORISATION TF-IDF

### 4.1 Méthode

TF-IDF (Term Frequency-Inverse Document Frequency)

### 4.2 Configuration Optimisée

- Vocabulaire limité à 5000 features les plus importantes
- Analyse des uni-grammes et bi-grammes
- Filtrage des termes trop rares ou trop fréquents
- Adaptation aux spécificités du français

---

## 5. ÉVALUATION COMPARATIVE DES MODÈLES

### 5.1 Modèles Testés

| Modèle | F1-Score (Test) | Validation Croisée | Avantages |
|---|---|---|---|
| **SVM** | **0.838** | **0.852 ± 0.018** | Efficace en haute dimension |
| Logistic Regression | 0.838 | 0.829 ± 0.011 | Rapide, interprétable |
| Random Forest | 0.832 | 0.829 ± 0.015 | Robuste, gestion non-linéarité |
| Gradient Boosting | 0.813 | 0.826 ± 0.013 | Ensemble, haute performance |
| Naive Bayes | 0.793 | 0.785 ± 0.003 | Probabiliste, rapide |

### 5.2 Résultat

**SVM sélectionné** comme modèle final basé sur le F1-Score pondéré et la stabilité en validation croisée.

---

## 6. CRÉATION DU PIPELINE FINAL

### 6.1 Architecture

```
Texte brut → Preprocessing français → Vectorisation TF-IDF → Modèle SVM → Prédiction
```

### 6.2 Modèle Retenu

#### Support Vector Machine (SVM)

- F1-Score sur test : 0.838
- Validation croisée : 0.852 ± 0.018
- Justification : Meilleure performance globale en validation croisée

### 6.3 Sauvegarde

- Format pickle avec toutes les étapes du pipeline
- Métadonnées incluant les performances et paramètres
- Horodatage pour le versioning

---

# 7. DÉVELOPPEMENT APPLICATION WEB

## 7.1 Interface Streamlit

Application web complète pour l'analyse de commentaires avec interface moderne et design responsive.

## 7.2 Trois Modes d'Utilisation

### Mode 1 : Analyse Simple

- Zone de saisie : Texte libre pour commentaires
- Prédiction instantanée : Classification en temps réel
- Visualisation : Graphiques de confiance
- Probabilités détaillées : Distribution par classe

### Mode 2 : Analyse de Masse

- Upload CSV : Glisser-déposer de fichiers
- Traitement par lots : Barre de progression
- Statistiques visuelles : Graphiques en secteurs
- Export résultats : CSV téléchargeable

### Mode 3 : Exemples Prédéfinis

- 5 types de commentaires : Positif, négatif, neutre, mitigé, enthousiaste
- Tests rapides : Démonstration des capacités
- Validation modèle : Vérification performance

---

# 8. OPTIMISATION ET DÉPLOIEMENT

## 8.1 Optimisations Techniques

- Cache intelligent pour le modèle
- Session state : Évite rechargements
- Gestion d'erreurs : Messages utilisateur clairs
- Compatibilité : Classes preprocessing incluses

## 8.2 Informations Modèle

**Sidebar informative** incluant :

- Métadonnées : F1-score, accuracy, taille dataset
- Classes détectées : Positif, neutre, négatif

- Guide utilisation : Instructions détaillées

---

# 9. RÉSULTATS ET PERFORMANCES

## 9.1 Métriques du Modèle Final

**Support Vector Machine (SVM)**

**Performances obtenues :**

- F1-Score pondéré (test) : 0.838
- Validation croisée : 0.852 ± 0.018
- Justification du choix : Meilleure stabilité en validation croisée

## 9.2 Comparaison des Modèles

- SVM et Logistic Regression : Performances similaires sur le test (0.838)
- SVM supérieur en validation croisée (0.852 vs 0.829)
- Naive Bayes : Performance moindre (0.793)
- Gradient Boosting : Performance intermédiaire (0.813)

## 9.3 Capacités Démontrées

- Détection positive : Reconnaissance satisfaction client
- Détection négative : Identification problèmes produit
- Nuances neutres : Commentaires mitigés
- Confiance mesurée : Probabilités par classe

---

# 10. TECHNOLOGIES UTILISÉES

## 10.1 Web Scraping

- Scrapy 2.8+
- Regex pour parsing
- Rotation user-agents

## 10.2 Machine Learning

- scikit-learn 1.3.0
- NLTK 3.8.1 (français)
- pandas, numpy

## 10.3 Interface Web

- Streamlit 1.28.0
- Plotly 5.17.0 (visualisations)
- CSS personnalisé

---

## 11. INNOVATION ET CONTRIBUTIONS

### 11.1 Aspects Novateurs

- Web scraping intelligent : Échantillonnage multi-catégories
- Preprocessing français avancé : Optimisé Amazon
- Pipeline complet : De la collecte au déploiement
- Interface utilisateur : Application web professionnelle

### 11.2 Défis Techniques Surmontés

- Anti-scraping Amazon : Contournement éthique
- Déséquilibre classes : Stratégie d'équilibrage optimisée
- Preprocessing français : Adaptation linguistique
- Déploiement modèle : Classes personnalisées

### 11.3 Applications Pratiques

- E-commerce : Monitoring satisfaction client
- Marketing : Analyse retours produits
- Support client : Détection problèmes automatique
- Business intelligence : Tendances des commentaires

---

## 12. RÉSULTATS OBTENUS

Ce projet présente un système d'analyse de commentaires pour les commentaires clients d'Amazon, développé à travers les étapes de collecte, preprocessing, modélisation et déploiement.

- Collecte de données Amazon
- Preprocessing spécialisé pour la langue française
- Évaluation comparative de modèles
- Application web avec Streamlit
- Documentation et code

---

*Projet de classification de commentaires utilisant web scraping, machine learning, NLP français et développement web.*