Université Abdelmalek Essaâdi

École Nationale des Sciences Appliquées de Tétouan

Filière : Big Data & IA

Module : Analyse du Web

Rapport du Projet : Analyse des tendances de mode à partir de données web à l'aide de l'intelligence artificielle

---

Réalisé par :

Hamza Kholti
Aya Salim
Anouar Bouzhar
Boutaina Bendaoud

Proposé par :

Mr. Imad Sassi

Année Universitaire : 2024-2025

Remerciements

Nous tenons à exprimer notre profonde gratitude envers toutes les personnes qui ont contribué, de près ou de loin, à la réalisation de ce projet de recherche.

Nos remerciements les plus sincères vont tout particulièrement au Professeur Imad Sassi pour son encadrement éclairé, ses conseils avisés et son soutien constant tout au long de cette étude.

Sa disponibilité, sa rigueur scientifique et son accompagnement bienveillant ont été des atouts précieux dans l'aboutissement de ce travail.

Table des matières

1 Contexte et Problématique **6**
- 1.1 L'Humanité à l'Ère de la Mode Digitale 6
- 1.2 Les Défis de la Compréhension Collective et l'Urgence Économique 6

2 Objectifs Stratégiques **7**
- 2.1 Vision Humaniste de l'Analyse Prédictive 7
- 2.2 Démocratiser l'Accès aux Insights Mode 7
- 2.3 Réconcilier Global et Local 7
- 2.4 Vers une Mode Plus Durable et Éthique 8

3 Architecture du Système **9**
- 3.1 Couche d'Ingestion des Données 9
- 3.2 Couche de Stockage des Données 10
- 3.3 Couche de Traitement des Données 11
- 3.4 Couche de Visualisation des Données 12

4 Collecte des Données **13**
- 4.1 Sources de Données 14
- 4.2 Méthodes de Collecte des Données 15
- 4.3 Stockage des Données dans MongoDB 16

5 Traitement des données collectées **18**
- 5.1 Type de données collectées 18
  - 5.1.1 Données visuelles (images) 18
  - 5.1.2 Données textuelles (commentaires et descriptions) 18
  - 5.1.3 Données numériques (métriques quantitatives) 18
- 5.2 Traitement des images 19
  - 5.2.1 Détection des vêtements avec YOLO 19
  - 5.2.2 Segmentation des vêtements avec SAM 19
  - 5.2.3 Encodage avec Autoencodeur 20
  - 5.2.4 Réduction de dimension avec PCA 20
  - 5.2.5 Détection de la couleur dominante dans les vêtements 20
- 5.3 Traitement de texte 21
  - 5.3.1 Analyse de sentiment 21
  - 5.3.2 Classification des descriptions eBay avec Sentence-BERT 22
- 5.4 Clustering 22
  - 5.4.1 Méthodologie utilisée 23
  - 5.4.2 Résultats du clustering 23

6 Modèles utilisés **23**
- 6.1 YOLO – Détection des vêtements 23
  - 6.1.1 Définition de YOLO 24
  - 6.1.2 Données d'entraînement utilisées 24
  - 6.1.3 Entraînement du modèle 24
  - 6.1.4 Analyse des performances 25

6.2 SAM – Segment Anything Model . . . 26
6.2.1 Définition de SAM . . . 26
6.2.2 Avantages observés . . . 26
6.2.3 Limites éventuelles . . . 26
6.3 Autoencodeur – Représentation visuelle des vêtements . . . 27
6.3.1 Analyse des résultats . . . 28
6.4 Analyse de sentiment . . . 28
6.4.1 Modèles testés . . . 28
6.4.2 Résultats obtenus . . . 28
6.4.3 Analyse comparative . . . 29

7 Dashboard Instagram . . . 29
7.1 Fonctionnement de dashboard : . . . 29
7.2 Les Filtres du Dashboard : . . . 30
7.3 Analyse et intérprétation des resultats . . . 31
7.3.1 Analyse des KPI Globaux . . . 31
7.3.2 Analyse des Visualisation : . . . 33

8 Dashboard eBay . . . 36
8.1 Fonctionnement du Dashboard : . . . 36
8.2 Les Filtres du Dashboard . . . 37
8.2.1 Filtre par Sexe . . . 37
8.2.2 Filtre par Sentiment . . . 37
8.2.3 Navigation Catégorielle Interactive . . . 38
8.3 Analyse des Résultats et Interprétation . . . 38
8.4 Analyse des KPI Globaux . . . 38
8.4.1 Volume Total de Produits (1000) . . . 38
8.4.2 Total Feedbacks (813,804) . . . 38
8.4.3 Prix Moyen Global (533,61$) . . . 38
8.5 Analyse des Visualisations . . . 38
8.5.1 Top Catégories par Feedback . . . 38
8.5.2 Distribution Géographique . . . 39
8.5.3 Tendances Couleurs . . . 40
8.5.4 Segmentation Prix . . . 41

9 Défis et Perspectives . . . 42
9.1 Défis rencontrés . . . 42
9.2 Perspectives d'évolution . . . 42

10 conclusion . . . 43

Liste des tableaux

1 Comparaison des modèles d'analyse de sentiment . . . 29

Table des figures

1 Architecture du système pour l'analyse des tendances de mode. 9
2 Flux de collecte des données pour l'analyse des tendances de mode, illustrant l'intégration des données provenant d'Instagram, Twitter, NewsAPI et eBay. 14
3 Architecture de stockage des données dans MongoDB, illustrant l'organisation des données brutes 17
4 Pipeline de traitement d'image 19
5 Analyse de sentiment 21
6 Pipeline de classification des description Ebay 22
7 Clustering avec HDBSCAN 22
8 Evaluation de Modele : YOLO 24
9 Prediction sur un image de test avec YOLO 25
10 Architecture Encodeur - Decodeur 27
11 Evaluation de performance de l'autoencodeur 27
12 Premiére page du dashboard d'Instagramm 30
13 Deuxiéme page du dashboard d'Instagramm 30
14 Somme de likes et Nombre de timestamp par Mois et page 33
15 Nombre de pages par cluster 34
16 Somme de Likes par owner 34
17 Répartition globale des sentiments par pages 35
18 Répartition globale des types de commentaires 36
19 Dashboard d'ebay 37
20 Top Catégories par Feedback 39
21 Nombre de catégory par location 40
22 Couleur tendances 41
23 Prix moyenne par catégorie 42

Introduction

La mode, miroir de nos sociétés et expression de nos identités, traverse aujourd'hui une révolution silencieuse mais profonde. Dans un monde hyperconnecté où chaque individu devient à la fois créateur et prescripteur de style, les codes traditionnels de l'industrie textile se réinventent quotidiennement. Les réseaux sociaux ont démocratisé l'influence, transformant chaque utilisateur en potentiel ambassadeur de tendances, tandis que les plateformes d'e-commerce révèlent en temps réel les aspirations et les choix de millions de consommateurs. Cette transformation digitale de l'univers vestimentaire soulève une question fondamentale : comment comprendre et anticiper les désirs d'une humanité qui s'exprime désormais massivement à travers ses choix esthétiques numériques ? Notre recherche propose une réponse innovante en exploitant la richesse des données générées spontanément par les utilisateurs d'Instagram et d'eBay. En analysant leurs sentiments, leurs préférences et leurs comportements d'achat, nous développons un système capable de déceler les tendances émergentes et de prédire l'évolution des goûts vestimentaires à l'échelle planétaire. Cette approche humaniste de l'analyse de données place l'individu au cœur de la compréhension des phénomènes de mode, reconnaissant que derrière chaque clic, chaque partage et chaque achat se cache une personne qui exprime sa personnalité et ses aspirations. Notre travail s'inscrit ainsi dans une démarche respectueuse de cette expression humaine tout en cherchant à en décoder les patterns collectifs pour mieux servir les besoins de chacun.

1 Contexte et Problématique

1.1 L'Humanité à l'Ère de la Mode Digitale

Nous vivons une époque fascinante où la frontière entre le réel et le virtuel s'estompe, particulièrement dans l'univers de la mode. Chaque jour, des millions d'individus partagent leurs tenues, leurs coups de cœur et leurs inspirations sur les réseaux sociaux, créant une gigantesque bibliothèque vivante du goût contemporain. Cette démocratisation de l'expression stylistique représente un phénomène anthropologique majeur : pour la première fois dans l'histoire de l'humanité, nous pouvons observer en temps réel les préférences esthétiques de populations entières.

Instagram est devenu le nouveau miroir de nos sociétés, où chaque selfie, chaque story, chaque post reflète non seulement un choix vestimentaire mais aussi une aspiration, une émotion, un message. Parallèlement, les plateformes comme eBay révèlent les comportements d'achat réels, traduisant les désirs en actions concrètes. Cette dualité entre l'aspiration (Instagram) et l'action (eBay) offre une fenêtre unique sur la psychologie collective de la consommation vestimentaire.

Cependant, cette richesse informationnelle soulève des défis éthiques et méthodologiques considérables. Comment analyser ces données tout en respectant l'intimité et l'authenticité des expressions individuelles ? Comment extraire des tendances collectives sans réduire la diversité des goûts personnels ? Ces questions guident notre approche, qui cherche à concilier performance analytique et respect de la dimension humaine de la mode.

1.2 Les Défis de la Compréhension Collective et l'Urgence Économique

L'abondance de données générées par les utilisateurs crée paradoxalement de nouvelles formes de complexité. Les marques et les créateurs, traditionnellement habitués à dicter les tendances depuis leurs ateliers parisiens ou milanais, se trouvent désormais face à un public qui s'auto-influence et crée ses propres codes esthétiques. Cette inversion du paradigme créatif bouscule les certitudes établies et exige de nouveaux outils de compréhension.

Aujourd'hui, les marques de mode perdent des millions d'euros chaque année parce qu'elles ratent les tendances ou produisent des vêtements que personne ne veut acheter. Cette réalité économique brutale illustre l'inadéquation croissante entre les méthodes traditionnelles de prédiction des tendances et la réalité mouvante des désirs des consommateurs. Les cycles de mode sont devenus ultra-rapides : ce qui est tendance aujourd'hui peut être dépassé demain.

La vitesse de propagation des tendances s'est également accélérée de manière exponentielle. Un motif, une couleur, une silhouette peuvent se diffuser instantanément à travers les continents, créant des phénomènes de mode globaux en quelques heures. Cette rapidité, si elle offre des opportunités inédites, rend également plus complexe l'anticipation des évolutions stylistiques. Les cycles traditionnels de six mois entre conception et commercialisation semblent désormais archaïques face à cette immédiateté numérique.

Par ailleurs, la fragmentation des sources d'influence complique la lecture des signaux.

Alors qu'autrefois quelques magazines et défilés suffisaient à comprendre les orientations de la mode, aujourd'hui l'influence se disperse entre millions de créateurs de contenu, chacun apportant sa pierre à l'édifice collectif du goût contemporain. Cette atomisation de l'influence, démocratique et enrichissante, nécessite de nouveaux instruments d'analyse pour saisir la symphonie complexe des préférences individuelles.

Notre défi ? Transformer le chaos apparent des réseaux sociaux en intelligence stratégique pour anticiper les goûts des consommateurs en temps réel.

2 Objectifs Stratégiques

2.1 Vision Humaniste de l'Analyse Prédictive

Notre projet s'articule autour d'une conviction profonde : derrière chaque donnée se cache une personne avec ses espoirs, ses goûts et ses rêves. L'objectif n'est pas de manipuler ou de formater les choix individuels, mais de mieux comprendre les aspirations collectives pour créer une mode plus inclusive, plus réactive et plus respectueuse des diversités culturelles et personnelles.

Nous aspirons à développer un écosystème intelligent capable de détecter les émergences créatives spontanées, de comprendre leurs résonances émotionnelles et de prédire leur potentiel de diffusion. Cette approche respecte l'authenticité des expressions individuelles tout en révélant les patterns collectifs qui façonnent nos préférences esthétiques communes.

2.2 Démocratiser l'Accès aux Insights Mode

L'un de nos objectifs fondamentaux consiste à démocratiser l'accès à l'intelligence tendance. Traditionnellement réservée aux grandes maisons de mode disposant de budgets conséquents pour leurs cabinets de style, la compréhension fine des évolutions esthétiques doit pouvoir bénéficier aux créateurs indépendants, aux petites marques et aux entrepreneurs de la mode.

En exploitant les données publiquement disponibles et en développant des outils accessibles, nous contribuons à une démocratisation de l'innovation dans l'industrie textile.

Cette démocratisation passe par la création d'un système capable d'identifier automatiquement les catégories vestimentaires en émergence, de décoder les palettes coloristiques porteuses d'avenir et d'analyser les sentiments associés aux différentes tendances. L'objectif est de fournir à tous les acteurs de la mode, indépendamment de leur taille, les clés de compréhension des aspirations contemporaines.

2.3 Réconcilier Global et Local

Dans un monde globalisé où les tendances se diffusent instantanément, nous observons paradoxalement un retour aux spécificités locales et culturelles. Notre système ambitionne de capturer cette dualité en identifiant les mouvements universels tout en respectant les particularismes régionaux.

L'analyse comparative entre différentes zones géographiques révèle comment une même

tendance peut se décliner différemment selon les contextes culturels, climatiques ou socio-économiques.

Cette approche géo-différenciée permet aux marques de développer des stratégies à la fois globalement cohérentes et localement pertinentes. Elle respecte la diversité des expressions culturelles tout en identifiant les points de convergence qui facilitent les économies d'échelle.

2.4 Vers une Mode Plus Durable et Éthique

Au-delà de l'efficacité commerciale, notre projet porte une ambition environnementale et sociale. En améliorant la précision des prédictions tendance, nous contribuons à réduire le gaspillage textile lié aux invendus. Une meilleure compréhension des aspirations réelles des consommateurs permet de produire plus juste, réduisant l'impact environnemental de l'industrie textile.

L'analyse des sentiments associés aux différentes pratiques de consommation révèle également l'évolution des consciences vers des choix plus responsables. Notre système peut ainsi accompagner la transition vers une mode plus éthique en identifiant les signaux précurseurs des changements de mentalité collective.

Cette approche humaniste et durable de l'analyse prédictive transforme la donnée en outil d'amélioration collective, servant à la fois les intérêts économiques des acteurs de la mode et les aspirations sociétales vers plus de responsabilité et d'authenticité.

3 Architecture du Système

L'architecture du système est conçue avec une précision rigoureuse pour offrir une analyse approfondie et dynamique des tendances de mode, permettant aux parties prenantes de découvrir des informations exploitables sur les préférences des consommateurs, les dynamiques du marché et les styles émergents. En intégrant harmonieusement l'ingestion, le stockage, le traitement et la visualisation des données, cette architecture forme un pipeline cohérent qui transforme des données brutes et hétérogènes en résultats stratégiques et exploitables. Construite pour garantir l'évolutivité, l'efficacité et la conformité aux réglementations sur la protection des données, telles que le RGPD, elle permet aux parties prenantes d'anticiper et de s'adapter aux évolutions rapides du paysage de la mode. L'architecture est illustrée dans la Figure 1, qui offre une vue d'ensemble visuelle du flux de données à travers chaque couche, clarifiant la structure opérationnelle du système.

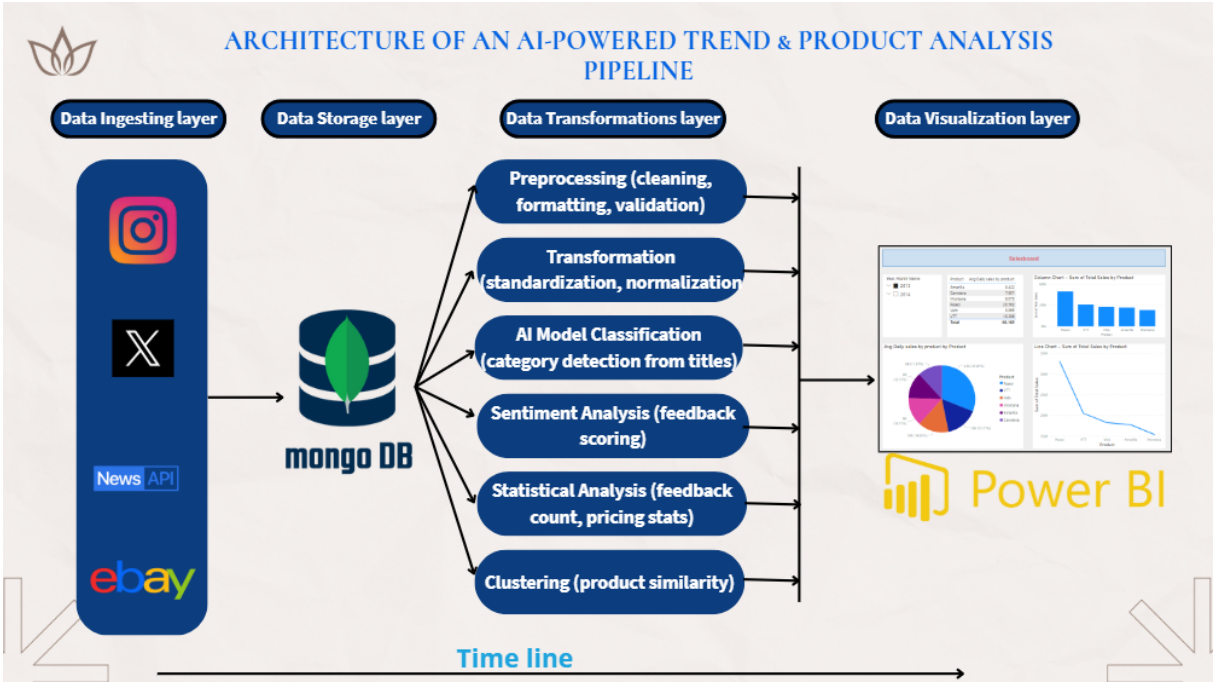

FIGURE 1 – Architecture du système pour l'analyse des tendances de mode.

3.1 Couche d'Ingestion des Données

La couche d'ingestion des données constitue le socle fondamental de l'architecture, orchestrant la collecte de données variées à partir de multiples plateformes pour capturer une vision globale des tendances de mode et du comportement des consommateurs. En ciblant un ensemble de variables soigneusement sélectionnées, cette couche garantit que le système recueille des données pertinentes, de haute qualité et adaptées à l'analyse des tendances. Chaque variable est choisie pour sa contribution unique à la compréhension des tendances, de l'engagement et des dynamiques du marché :

- **Métadonnées des publications** : Incluent les identifiants des publications et les attributs spécifiques à la plateforme (par exemple, type de publication, plateforme source), garantissant un suivi précis et une différenciation des contenus, essentiels pour maintenir l'intégrité des données et faciliter leur intégration.

- **Horodatages** : Enregistrent la date et l'heure des publications ou des annonces, permettant d'analyser les motifs temporels, d'identifier les variations saisonnières et de suivre le cycle de vie des tendances, crucial pour comprendre les dynamiques temporelles de la mode.
- **Mentions et mots-clés de tendances** : Capturent les références à des marques, créateurs ou styles (par exemple, "mode durable", "Gucci") dans le texte, mettant en lumière les entités influentes et les tendances émergentes qui façonnent les préférences des consommateurs.
- **Hashtags** : Étiquettes (par exemple, minimalisme, streetwear) servant à catégoriser les publications, facilitant l'identification des catégories de mode tendance et des regroupements thématiques reflétant les mouvements populaires.
- **Commentaires** : Fournissent des réponses textuelles générées par les utilisateurs, offrant une source riche d'informations qualitatives sur le sentiment des consommateurs, leurs préférences et leurs réactions aux contenus de mode.
- **Nombre de likes** : Mesure l'engagement des publications, reflétant la popularité et l'impact social des articles ou styles de mode, permettant d'évaluer leur résonance auprès des communautés.
- **Titres et prix des produits** : Incluent les noms des articles et leurs coûts sur les plateformes de vente, permettant d'analyser les tendances dictées par le marché, les comportements de dépenses des consommateurs et les stratégies de tarification.
- **Retours/commentaires** : Capturent les avis des acheteurs, fournissant des informations sur la réception des produits, les perceptions de qualité et la satisfaction des clients, clés pour évaluer l'adoption des tendances.
- **Réputation des vendeurs** : Reflète les évaluations ou métriques de crédibilité des vendeurs, aidant à évaluer la confiance, qui influence les décisions d'achat des consommateurs.
- **Prix et type de livraison** : Englobent les coûts et méthodes de livraison (par exemple, standard, express), révélant les facteurs logistiques impactant le comportement d'achat et l'accessibilité du marché.
- **URLs d'images** : Fournissent des liens vers des contenus visuels, permettant l'analyse des couleurs, motifs et styles pour identifier les tendances esthétiques.

La collecte des données est réalisée à l'aide d'outils basés sur Python, avec des processus de validation rigoureux pour garantir la qualité, la cohérence et le respect des politiques des plateformes. Les détails des sources de données seront approfondis dans les sections suivantes.

3.2 Couche de Stockage des Données

La couche de stockage des données est conçue pour gérer efficacement les ensembles de données vastes et variés collectés, garantissant leur stockage sécurisé, leur accessibilité et leur préparation pour l'analyse. MongoDB est choisi comme solution de stockage en raison de sa flexibilité exceptionnelle pour gérer des données non structurées et semi-structurées à travers les variables ciblées. La base de données héberge à la fois les données brutes (par exemple, commentaires originaux, URLs d'images) et les résultats traités (par exemple, prix standardisés, scores de sentiment), avec des processus de normalisation et d'anonymisation appliqués pour assurer la conformité aux réglementations de protection des données

comme le RGPD. La conception évolutive et les capacités d'indexation efficaces de MongoDB permettent un stockage et une récupération rapides, répondant aux exigences de l'analyse des tendances à grande échelle tout en maintenant l'intégrité et la sécurité des données.

3.3 Couche de Traitement des Données

La couche de traitement des données représente le cœur analytique du système, transformant les données brutes et hétérogènes en informations exploitables grâce à une série d'étapes soigneusement conçues et adaptées aux variables ciblées. Cette couche exploite des techniques avancées pour nettoyer, analyser et modéliser les données, révélant des motifs qui mettent en lumière les nuances des tendances de mode et du comportement des consommateurs :

- **Nettoyage et Standardisation des Données** : Garantit la cohérence et l'utilité des données en éliminant les informations non pertinentes et en standardisant les formats pour une analyse fluide. Les variables inutiles, telles que les champs génériques de type de données (par exemple, "texte" ou "numérique"), sont supprimées et remplacées par des valeurs inférées à partir de variables pertinentes comme les prix et types de livraison (par exemple, inférer une livraison express à partir de coûts élevés). Les valeurs manquantes sont gérées avec soin (par exemple, imputation de prix moyens pour les entrées incomplètes), et les formats de texte sont normalisés (par exemple, uniformisation des orthographes de hashtags comme StreetWear en streetwear, harmonisation des formats de devises pour les prix). Les fourchettes de prix (par exemple, "de 3040") sont converties en valeurs uniques (par exemple, 35 en calculant la moyenne), et les formats de retours (par exemple, "feedback(152,36)" en 152,36) sont standardisés pour assurer une analyse cohérente. Ces étapes rigoureuses créent un ensemble de données propre, unifié et fiable, essentiel pour les analyses avancées et la génération d'insights précis.
- **Extraction de Sentiments** : Utilise des modèles basés sur BERT de Hugging Face pour analyser les commentaires et les retours, classifiant les sentiments des consommateurs (par exemple, positif, négatif, neutre) et attribuant des scores de polarité. Cette analyse approfondie permet de révéler l'enthousiasme, l'insatisfaction ou l'indifférence des consommateurs face aux articles ou tendances de mode, priorisant les tendances qui captent fortement l'attention du public et fournissant des informations clés pour orienter les stratégies marketing et de développement de produits.
- **Méthodes de Regroupement** : Regroupe les hashtags, mots-clés de tendances et titres de produits en clusters thématiques (par exemple, vintage, athleisure) et aligne les comportements similaires des consommateurs (par exemple, acheteurs éco-conscients, amateurs de luxe) pour identifier des catégories de mode cohérentes et des segments d'audience. Ce processus permet une compréhension nuancée de la cohésion des tendances et des préférences des consommateurs, facilitant une segmentation ciblée et une personnalisation des stratégies de marché.
- **Analyse des Couleurs et Motifs** : Exploite YOLO pour détecter et extraire les couleurs dominantes (par exemple, bleu pastel, rouge vif) et les motifs (par exemple, floral, géométrique) à partir des URLs d'images, révélant les tendances esthétiques qui définissent les préférences visuelles des consommateurs. Cette analyse visuelle est cruciale pour identifier les éléments stylistiques qui influencent les directions du

marché et les choix des créateurs.

- **Méthodes Statistiques** : Applique des analyses de corrélation et de dépendance pour découvrir des relations significatives entre les variables, comme entre le nombre de likes et les mots-clés de tendances ou entre les prix et la réputation des vendeurs. Des tests de signification des tendances sont également effectués pour valider l'importance et la fiabilité des tendances identifiées, garantissant que les insights sont robustes et exploitables pour une prise de décision stratégique.
- **Modèles de Segmentation** : Mesure la similarité des articles de mode en segmentant les produits et les consommateurs selon des caractéristiques comme le style, la couleur, le prix ou le sentiment, identifiant des tendances de niche et des segments de consommateurs distincts (par exemple, acheteurs à budget limité vs acheteurs de luxe). Cette segmentation fine permet de cibler des niches de marché spécifiques avec une précision accrue, renforçant l'efficacité des stratégies marketing.
- **Catégorisation** : Attribue les publications et les produits à des taxonomies standardisées de l'industrie de la mode (par exemple, décontracté, formel, durable) en utilisant des modèles Random Forest, en s'appuyant sur les hashtags, mots-clés et caractéristiques visuelles. Cette catégorisation structurée aligne les tendances avec les normes du marché, facilitant des insights exploitables pour les parties prenantes et renforçant l'alignement stratégique avec les attentes de l'industrie.

Ces processus, mis en œuvre à l'aide de bibliothèques Python telles que `Pandas`, `Scikit-learn`, les modèles BERT de Hugging Face pour l'analyse textuelle et YOLO pour l'analyse d'images, fournissent des insights complets sur les motifs temporels, thématiques et de sentiment, constituant le socle de l'analyse des tendances de mode.

3.4 Couche de Visualisation des Données

La couche de visualisation des données est conçue pour transformer des insights analytiques complexes en représentations visuelles accessibles, intuitives et exploitables, permettant aux parties prenantes d'explorer les tendances, d'identifier les opportunités et de prendre des décisions basées sur les données avec confiance. Les objectifs principaux de cette couche sont :

- **Identification des Tendances** : Mettre en évidence les tendances dominantes et émergentes de la mode à travers des visualisations claires et intuitives des hashtags, mots-clés et éléments visuels, permettant aux parties prenantes de saisir rapidement les directions du marché et d'identifier les opportunités stratégiques.
- **Analyse Temporelle** : Visualiser l'évolution des tendances au fil du temps, révélant les motifs saisonniers, les cycles de popularité et les signaux prédictifs pour anticiper les futurs changements, guidant ainsi la planification stratégique des entreprises.
- **Exploration Interactive** : Fournir des outils dynamiques permettant aux parties prenantes de filtrer et d'explorer des variables spécifiques, des thèmes ou des périodes temporelles, offrant une flexibilité pour répondre à divers besoins professionnels, qu'il s'agisse de marketing, de design ou de gestion des stocks.
- **Insights sur l'Engagement** : Afficher des métriques d'engagement telles que le nombre de likes et les scores de sentiment pour évaluer l'enthousiasme des consommateurs, leurs préférences et la réception du marché, informant ainsi les stratégies de marketing, de positionnement de produits et de développement de marque.

Ces insights sont fournis via un tableau de bord interactif Power BI, proposant des types de graphiques soigneusement adaptés aux variables ciblées pour maximiser la clarté et l'impact :

- **Graphiques en Ligne** : Représentent la fréquence des hashtags ou mots-clés au fil du temps, en utilisant les horodatages, pour visualiser les cycles de vie des tendances et identifier les pics saisonniers, comme l'essor de modeDurable au printemps, offrant une vue claire des dynamiques temporelles.
- **Histogrammes** : Comparent le nombre de likes ou les prix des produits entre catégories ou périodes temporelles, fournissant des insights sur les niveaux d'engagement des consommateurs et les tendances de tarification qui façonnent les comportements d'achat.
- **Cartes de Chaleur** : Cartographient les tendances de couleurs extraites des URLs d'images en fonction des horodatages, révélant les préférences esthétiques saisonnières, comme la dominance des tons pastel dans les collections d'été, pour guider les décisions de design.
- **Filtres Interactifs** : Permettent aux parties prenantes d'explorer les données par variable (par exemple, réputation des vendeurs, type de livraison), période temporelle ou thème, soutenant une analyse dynamique et orientée utilisateur pour répondre aux besoins stratégiques spécifiques.

Les capacités de visualisation robustes de Power BI garantissent une interaction intuitive avec des ensembles de données complexes, permettant aux parties prenantes de naviguer facilement dans les insights et de découvrir des opportunités exploitables avec une clarté exceptionnelle.

4 Collecte des Données

La collecte des données représente une étape cruciale pour alimenter l'analyse des tendances de mode, en fournissant un socle d'informations brutes indispensable pour décrypter les préférences des consommateurs, les dynamiques du marché et les styles émergents. Cette phase vise à rassembler un ensemble diversifié et complémentaire de données à partir de sources stratégiquement choisies, chacune offrant une perspective unique sur les tendances et les comportements. Les variables ciblées englobent les métadonnées des publications, les horodatages, les mentions et mots-clés de tendances, les hashtags, les commentaires, le nombre de likes, les titres et prix des produits, les retours/commentaires, la réputation des vendeurs, les prix et types de livraison, ainsi que les URLs d'images. En s'appuyant sur des plateformes sociales, des sources médiatiques et des marchés en ligne, la collecte est orchestrée pour garantir la qualité, la pertinence et la conformité aux réglementations, notamment le RGPD. Le flux de collecte des données est illustré dans la Figure 3, qui présente une vue d'ensemble des sources et de leur intégration dans le système.

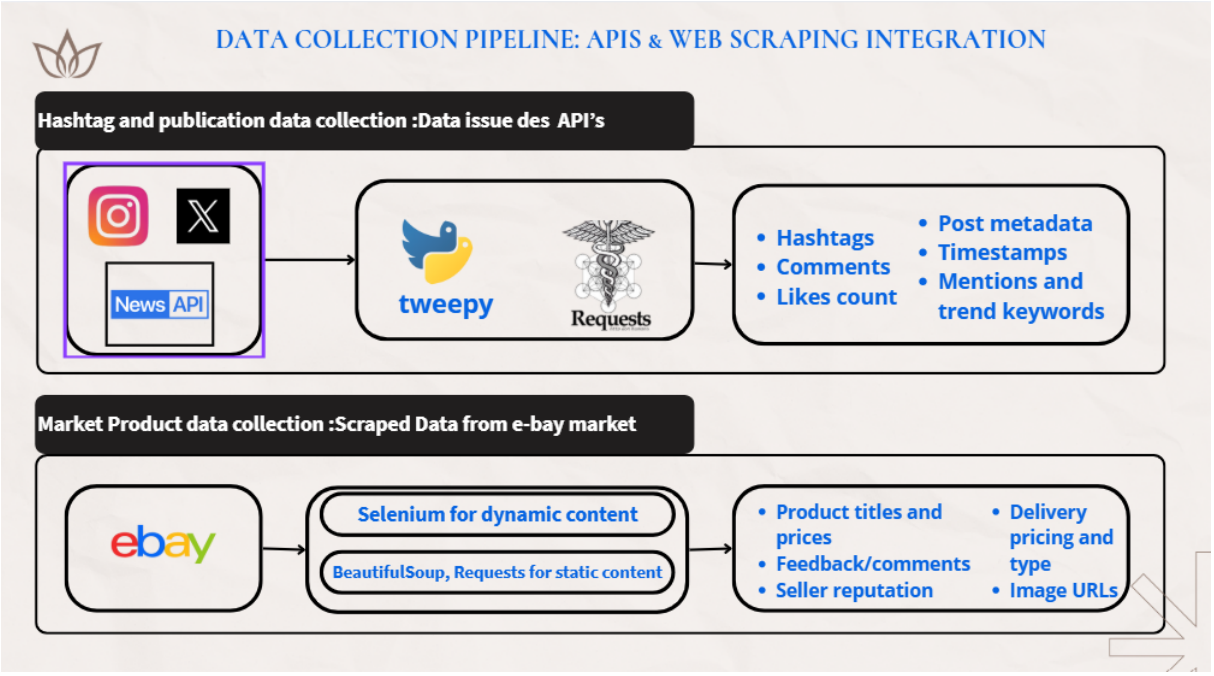

FIGURE 2 – Flux de collecte des données pour l'analyse des tendances de mode, illustrant l'intégration des données provenant d'Instagram, Twitter, NewsAPI et eBay.

4.1 Sources de Données

La collecte repose sur quatre sources principales, soigneusement sélectionnées pour leur capacité à fournir des données riches et complémentaires, enrichissant ainsi l'analyse des tendances de mode :

- **Instagram** : Une plateforme sociale incontournable pour explorer les tendances visuelles et sociales de la mode. Instagram offre des données variées, incluant les hashtags (par exemple, streetwear, modeDurable), les métadonnées des publications, les horodatages, les commentaires, le nombre de likes et les URLs d'images. Ces informations permettent de capturer les styles en vogue, les préférences esthétiques des utilisateurs et l'engagement communautaire, offrant une fenêtre directe sur les tendances émergentes et les influences culturelles qui façonnent le paysage de la mode.
- **Twitter** : Une source dynamique pour saisir les discussions en temps réel et les tendances sociales. Twitter fournit des données telles que les hashtags, les mentions et mots-clés de tendances, les métadonnées des publications, les horodatages et les commentaires. Cette plateforme permet de détecter les conversations spontanées autour des marques, des styles ou des événements de mode, révélant les sentiments des consommateurs et les sujets qui captent l'attention du public à un instant donné, enrichissant ainsi l'analyse avec des perspectives actuelles.
- **NewsAPI** : Une source précieuse pour accéder aux articles d'actualité liés à la mode, permettant de recueillir des mentions et mots-clés de tendances à partir des titres, résumés et contenus d'articles. NewsAPI apporte un contexte médiatique en mettant en lumière les tendances relayées par les médias, telles que les lancements de collections, les collaborations de marques ou les mouvements stylistiques comme

la mode durable. Ces données complètent les insights sociaux en offrant une vision élargie des influences culturelles et industrielles.

- **eBay** : Une plateforme de commerce en ligne essentielle pour explorer les données du marché. eBay fournit des informations détaillées sur les titres des produits, leurs prix, les retours et commentaires des acheteurs, la réputation des vendeurs, ainsi que les prix et types de livraison. Ces données permettent d'analyser les tendances d'achat, les préférences des consommateurs pour certains styles ou gammes de prix, et les facteurs logistiques influençant les décisions d'achat, offrant une perspective tangible des dynamiques commerciales dans le secteur de la mode.

4.2 Méthodes de Collecte des Données

Le processus de collecte repose sur deux approches complémentaires : l'ingestion via API pour les plateformes sociales et médiatiques, et le scraping pour les données de marché. Chaque méthode est adaptée aux spécificités des sources et aux variables ciblées.

- **Ingestion via API** : Cette méthode est utilisée pour collecter des données à partir d'Instagram, Twitter et NewsAPI, exploitant leurs interfaces de programmation officielles pour accéder aux données de manière structurée et efficace.
  - **Instagram** : L'API Graph d'Instagram permet de recueillir des données sociales riches, incluant les hashtags (par exemple, streetwear, modeDurable), les métadonnées des publications (identifiants, type de contenu), les horodatages, les commentaires, le nombre de likes et les URLs d'images. Ces données capturent les tendances visuelles, l'engagement des utilisateurs et les préférences esthétiques, offrant une perspective directe sur les styles et influences populaires dans la mode.
  - **Twitter** : L'API de Twitter est utilisée pour extraire les hashtags, les mentions et mots-clés de tendances, ainsi que les métadonnées des publications, les horodatages et les commentaires. Cette approche permet de saisir les discussions en temps réel sur les marques, les styles ou les événements de mode, révélant les sentiments des consommateurs et les sujets d'actualité qui façonnent les conversations.
  - **NewsAPI** : NewsAPI fournit un accès aux articles d'actualité liés à la mode, permettant de collecter des mentions et mots-clés de tendances à partir des titres, résumés et contenus d'articles. Ces données enrichissent l'analyse en capturant les tendances médiatisées, telles que les lancements de collections ou les mouvements stylistiques (par exemple, la mode durable), offrant un contexte culturel et industriel complémentaire.
- **Scraping des données d'eBay** : Pour collecter les données de marché, eBay est exploré à l'aide de techniques de scraping adaptées à la structure de la plateforme. Ce processus se déroule en deux étapes principales :
  - **Scraping des pages statiques avec 'requests' et 'BeautifulSoup'** : La bibliothèque Python 'requests' est utilisée pour accéder aux pages statiques d'eBay, telles que les listes de produits, tandis que 'BeautifulSoup' permet d'analyser le code HTML pour extraire les informations pertinentes. Cette étape consiste à collecter les liens vers les pages de chaque article, en identifiant les titres des produits, les prix initiaux et les métadonnées associées

(par exemple, catégories de produits). Cette approche permet de constituer un répertoire structuré des articles disponibles pour une analyse approfondie.

– **Scraping des données dynamiques avec 'Selenium'** : Pour les pages générées dynamiquement par du code JavaScript, comme les pages détaillées des articles, 'Selenium' est employé pour naviguer automatiquement vers chaque lien d'article extrait. Cette méthode permet de collecter des informations détaillées, incluant les prix finaux, les retours et commentaires des acheteurs, la réputation des vendeurs (par exemple, scores ou évaluations), les prix et types de livraison (standard, express), ainsi que les URLs d'images des produits. Ces données offrent une vision complète des dynamiques de marché, des préférences des consommateurs et des aspects logistiques influençant les achats.

4.3 Stockage des Données dans MongoDB

Le stockage des données est orchestré à l'aide de **MongoDB**, une base de données NoSQL choisie pour sa capacité à gérer de manière flexible les ensembles de données non structurées et semi-structurées générés par la collecte. Cette solution est idéale pour accueillir les variables ciblées, telles que les **métadonnées des publications**, les **horodatages**, les **mentions et mots-clés de tendances**, les **hashtags**, les **commentaires**, le **nombre de likes**, les **titres et prix des produits**, les **retours/commentaires**, la **réputation des vendeurs**, les **prix et types de livraison**, et les **URLs d'images**. MongoDB stocke à la fois les données brutes (par exemple, commentaires originaux, URLs d'images) et les données traitées (par exemple, prix standardisés, scores de sentiment), permettant une gestion centralisée et efficace. Les données sont normalisées pour assurer leur cohérence et anonymisées pour respecter les exigences du RGPD, garantissant la protection des informations personnelles. La conception évolutive de MongoDB, avec ses capacités d'indexation performantes, soutient l'analyse à grande échelle, offrant un accès rapide et fiable pour les étapes ultérieures de traitement et d'analyse des tendances de mode.

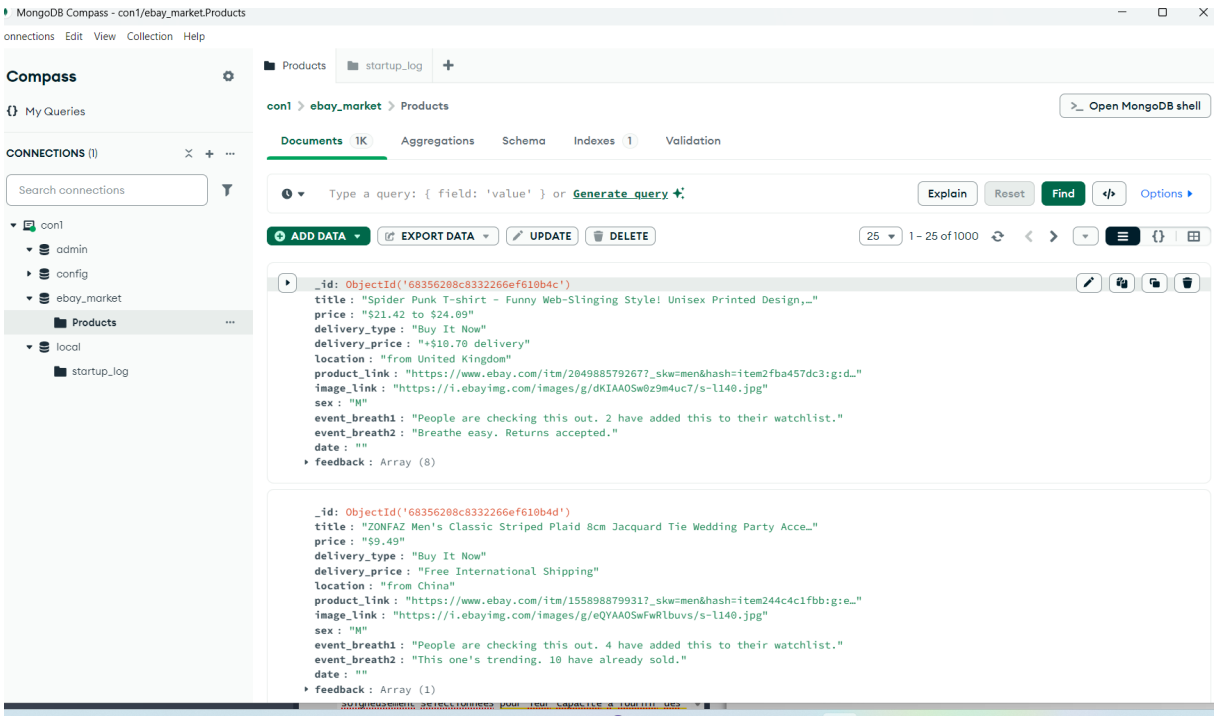

FIGURE 3 – Architecture de stockage des données dans MongoDB, illustrant l'organisation des données brutes

L'architecture du système et la collecte des données forment un pipeline robuste pour analyser les tendances de mode, transformant des données brutes en insights exploitables sur les préférences des consommateurs et les dynamiques du marché. L'architecture s'articule autour de quatre couches : l'**ingestion** collecte des données variées (métadonnées, hashtags, prix, images) à partir de sources comme Instagram, Twitter, NewsAPI et eBay ; le **stockage** utilise **MongoDB** pour gérer efficacement les données brutes et traitées, avec normalisation et anonymisation conformes au RGPD ; le **traitement** nettoie et analyse les données via des outils comme **BERT** pour les sentiments, **YOLO** pour les couleurs et motifs, et **Random Forest** pour la catégorisation, révélant des motifs thématiques et temporels ; la **visualisation** présente les insights via un tableau de bord interactif **Power BI**, avec des graphiques (lignes, nuages de mots, histogrammes, cartes de chaleur) pour explorer les tendances et l'engagement. La collecte des données s'appuie sur l'**ingestion via API** pour Instagram, Twitter et NewsAPI, capturant hashtags et publications, et sur le **scraping** d'eBay avec `requests` et `BeautifulSoup` pour les pages statiques et `Selenium` pour les données dynamiques, extrayant titres, prix, retours et détails logistiques. Ce cadre intégré garantit une analyse précise et évolutive des tendances de mode, soutenant des décisions stratégiques alignées sur les attentes du marché.

5 Traitement des données collectées

5.1 Type de données collectées

Dans notre projet d'analyse des tendances de la mode à partir de contenus en ligne, nous avons collecté un ensemble de données issues de deux plateformes principales : **Instagram** et **eBay**. Ces données sont variées à la fois dans leur nature (image, texte, valeur numérique) et dans leur rôle analytique (description, opinion, statistique).

5.1.1 Données visuelles (images)

Chaque post ou article possède son/ses image(s). Nous avons d'abord collecté un grand nombre d'**images** : Des **images de posts Instagram** qui montrent des personnes portant des vêtements, souvent dans un contexte naturel ou esthétique. Des **images d'articles eBay** représentant les vêtements à vendre, généralement sur fond blanc ou neutre, dans un style de présentation standardisé.

Ces images ont servi à détecter, segmenter, encoder et analyser les **vêtements visibles**, afin d'en extraire des **informations visuelles pertinentes** (comme la couleur dominante, les formes, et le style vestimentaire).

5.1.2 Données textuelles (commentaires et descriptions)

Nous avons ensuite collecté des **données textuelles** :

- Des **commentaires d'utilisateurs Instagram**, souvent courts, contenant des appréciations ou des réactions au post (positives, négatives ou neutres).
- Des **descriptions d'articles eBay**, souvent longues et riches en détails, mais sans structure prédéfinie (ni catégorie explicite).
- Des **feedbacks clients**, qui permettent d'évaluer la satisfaction des utilisateurs par rapport au produit.

Ces textes ont été analysés pour deux objectifs principaux :

- **Détecter le sentiment exprimé** dans les commentaires et feedbacks (positif, négatif, neutre).
- **Classer automatiquement les descriptions** des articles eBay dans des catégories vestimentaires (T-shirt, pantalon, veste, etc.), grâce à des modèles d'apprentissage du langage.

5.1.3 Données numériques (métriques quantitatives)

Enfin, nous avons collecté plusieurs **données numériques**, c'est-à-dire des valeurs chiffrées associées à chaque post ou produit. Cela inclut :

- Le **nombre de commentaires**, de réactions ou de likes sur un post Instagram.
- Le **nombre de feedbacks clients** sur un article eBay.
- Le **prix de vente** des articles, ainsi que les **frais de livraison** associés.
- Les **scores de sentiment** générés automatiquement par nos modèles de classification.
- Et d'autres indicateurs tels que la localisation du vendeur ou le type de livraison.

Ces données numériques ont joué un rôle essentiel pour la **quantification des tendances**, la construction de **KPI (indicateurs de performance)**, et pour alimenter nos algorithmes de **clustering**, afin de regrouper les articles similaires.

5.2 Traitement des images

FIGURE 4 – Pipeline de traitement d'image

Le traitement des images représente une étape centrale dans notre pipeline d'analyse, car il permet d'extraire automatiquement des **informations visuelles pertinentes** à partir des photos issues d'Instagram et des annonces eBay. L'objectif principal est d'identifier, d'isoler, puis de représenter de manière numérique les vêtements présents dans chaque image.

Pour cela, nous avons mis en place une **chaîne de traitement automatique** en plusieurs étapes, en combinant différents modèles de machine learning, chacun spécialisé dans une tâche précise.

5.2.1 Détection des vêtements avec YOLO

La première étape consiste à **localiser les vêtements dans l'image**. Nous utilisons pour cela le modèle **YOLOv8 (You Only Look Once)**, un réseau de neurones convolutifs spécialisé dans la **détection d'objets en temps réel**.

YOLO analyse l'image entière en un seul passage et retourne une ou plusieurs **zones d'intérêt** (bounding boxes) autour des objets détectés. Dans notre cas, ces objets sont principalement des vêtements comme : des t-shirts, des vestes et des pantalons. Cela nous permet de nous concentrer uniquement sur la zone utile de l'image.

5.2.2 Segmentation des vêtements avec SAM

Une fois les vêtements localisés, l'image et les boîtes englobantes détectées sont transmises au modèle **SAM (Segment Anything Model)**, développé par Meta.

SAM est un modèle de **segmentation automatique**, capable de produire un **masque précis** autour de l'objet situé dans une région d'intérêt. Contrairement à YOLO qui

détecte une zone rectangulaire, SAM permet de **distinguer le vêtement de son arrière-plan** pixel par pixel, ce qui est essentiel pour une analyse visuelle de qualité.

Grâce à la combinaison YOLO + SAM, nous obtenons des **découpes nettes et ciblées** des vêtements, sans bruit visuel.

5.2.3 Encodage avec Autoencodeur

Les vêtements segmentés sont ensuite passés à travers un **autoencodeur**, un type de réseau de neurones conçu pour apprendre une **représentation compressée** de données visuelles.

- L'**encodeur** transforme l'image segmentée du vêtement en un **vecteur latent**, qui est une version numérique réduite contenant les **informations visuelles les plus importantes** (formes, couleurs, textures...).
- L'**objectif principal** ici est de disposer d'une représentation compacte et exploitable de chaque vêtement, que l'on pourra utiliser pour du **clustering** ou des **analyses comparatives**.

5.2.4 Réduction de dimension avec PCA

Les vecteurs latents extraits par l'autoencodeur peuvent encore être un peu trop volumineux pour certaines analyses statistiques.

C'est pourquoi nous appliquons ensuite une **réduction de dimension** à l'aide de **PCA (Analyse en Composantes Principales)**. PCA permet de conserver uniquement les **dimensions les plus informatives** du vecteur, tout en supprimant le bruit, ce qui améliore l'efficacité du **clustering** et des visualisations, tout en réduisant le coût de calcul.

5.2.5 Détection de la couleur dominante dans les vêtements

La **couleur des vêtements** est un élément fondamental dans l'analyse des tendances de mode. Elle influence fortement les préférences des utilisateurs et peut être utilisée pour regrouper des articles similaires ou suivre l'évolution des styles dans le temps. Dans notre projet, nous avons donc intégré une étape dédiée à la **détection automatique de la couleur dominante** présente sur chaque vêtement.

Pour garantir une détection fiable et cohérente, l'analyse de la couleur ne se fait pas sur l'image entière, mais uniquement sur la **zone centrale du vêtement segmenté**, là où la couleur est la plus représentative et la moins affectée par le décor ou les ombres. Cette approche réduit les interférences dues à l'arrière-plan ou à d'autres objets présents dans l'image.

La couleur est ensuite extraite par une méthode de **clustering non supervisé**, qui permet d'identifier les différentes teintes présentes dans l'image, puis de déterminer laquelle est la plus dominante. Cette étape permet de résumer chaque vêtement par **une seule couleur principale**, exprimée sous forme de **nom de couleur standardisé** (comme *lightgray*, *navy*, *burgundy*, etc.), selon les définitions CSS3.

Cette information est ensuite utilisée pour enrichir la base de données produit. Elle nous permet : d'**afficher la couleur principale** de chaque vêtement dans le tableau de données, de **filtrer ou regrouper les articles par couleur dominante**, et d'**intégrer la couleur dans le processus de clustering**, en complément des autres caractéristiques visuelles et textuelles.

Grâce à cette méthode, la couleur devient un critère d'analyse exploitable à grande échelle, permettant de mieux **visualiser les tendances chromatiques** de la mode dans les posts Instagram ou les articles eBay.

5.3 Traitement de texte

En parallèle du traitement d'images, une partie essentielle de notre projet repose sur l'analyse des **textes associés aux publications**. Ces textes incluent notamment les **commentaires Instagram**, les **descriptions d'articles eBay**, ainsi que les **avis ou feedbacks utilisateurs**. Notre objectif est d'en extraire des **informations sémantiques et émotionnelles** utiles à l'analyse des tendances de la mode.

Nous avons structuré le traitement des textes autour de deux tâches principales : **l'analyse de sentiment** et **la classification sémantique des descriptions**.

5.3.1 Analyse de sentiment

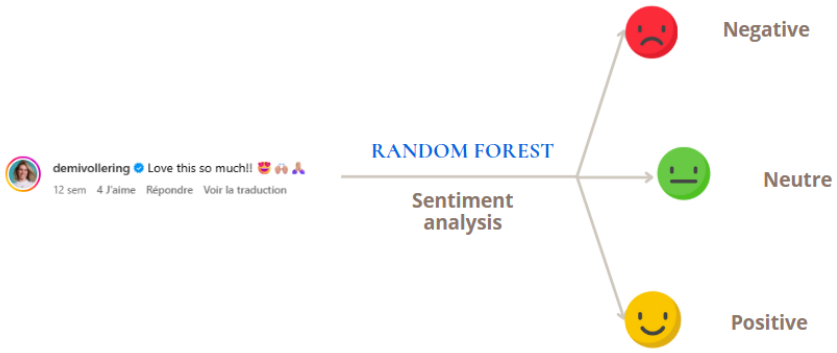

FIGURE 5 – Analyse de sentiment

Les commentaires des utilisateurs contiennent souvent des **indices émotionnels** sur leur opinion à propos d'un vêtement ou d'un article. Pour en extraire le **sentiment exprimé** (positif, négatif, neutre), nous avons expérimenté plusieurs modèles de classification.

- Nous avons d'abord testé des modèles traditionnels comme **Random Forest**, **XGBoost** et **K-Nearest Neighbors (KNN)**.
- Ces modèles ont été entraînés à partir de représentations vectorielles des textes, obtenues via **CountVectorizer** et **TfidfVectorizer**, deux méthodes classiques de transformation de texte en vecteurs numériques.
- Ensuite, pour améliorer la précision, nous avons utilisé un modèle de traitement du langage naturel avancé : **BERT (Bidirectional Encoder Representations from Transformers)** adapté à l'analyse de sentiments.

5.3.2 Classification des descriptions eBay avec Sentence-BERT

FIGURE 6 – Pipeline de classification des description Ebay

Les descriptions d'articles sur eBay ne mentionnent pas explicitement la **catégorie du vêtement** (ex. : T-shirt, pantalon, veste). Pour pallier cette absence, nous avons utilisé **Sentence-BERT**, une variante de BERT conçue pour produire des **représentations vectorielles efficaces de phrases entières**.

- Chaque description est encodée en un vecteur via **Sentence-BERT**.
- Ce vecteur est ensuite comparé à des vecteurs de **catégories prédéfinies** (ex. : T-shirt, chemise, robe...) à l'aide d'une **mesure de similarité** comme la cosinus-similarité.
- Le système attribue alors à chaque article la **catégorie la plus proche sémantiquement**.

Cette méthode nous a permis de reconstituer automatiquement une **catégorisation manquante** des produits eBay, essentielle pour les analyses de tendance par type de vêtement.

5.4 Clustering

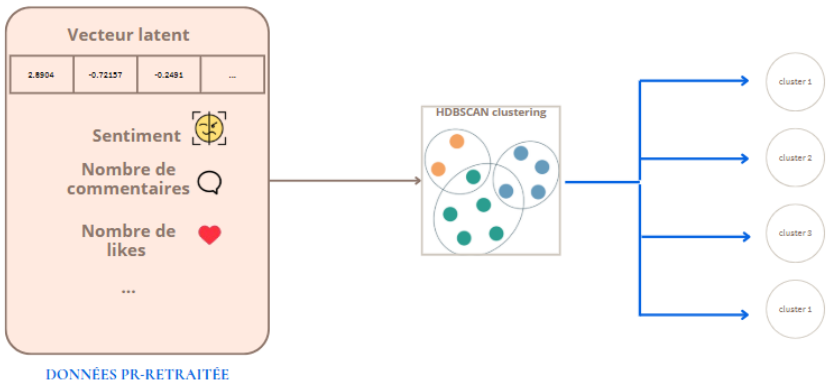

FIGURE 7 – Clustering avec HDBSCAN

Une fois les données visuelles et textuelles traitées et converties en représentations numériques exploitables (vecteurs latents, scores de sentiment, catégories, etc.), nous avons

appliqué une étape de **clustering**. L'objectif principal est de **regrouper automatiquement les articles ou publications similaires**, afin d'identifier des **groupes de tendances**, ou encore des **styles vestimentaires proches**.

Le clustering permet de répondre à plusieurs questions fondamentales : Quels articles se ressemblent visuellement (formes, couleurs) ?, Quels posts suscitent des sentiments similaires de la part des utilisateurs ? et Existe-t-il des groupes de vêtements associés à des niveaux d'intérêt proches (likes, commentaires, réactions) ?

En créant ces regroupements, nous sommes capables d'explorer les **tendances dominantes**, de découvrir des **patterns de mode** récurrents, et de poser les bases de **recommandations automatiques personnalisées**.

5.4.1 Méthodologie utilisée

Pour effectuer le clustering, nous avons utilisé une méthode non supervisée puissante : le **clustering hiérarchique HDBSCAN** (Hierarchical Density-Based Spatial Clustering of Applications with Noise), une version améliorée de l'algorithme DBSCAN.

Les données utilisées comme entrée pour ce modèle incluent : Les **vecteurs latents visuels** issus de l'autoencodeur (compressés par PCA), Les **catégories de vêtements**, Et dans certains cas, des **indicateurs d'engagement utilisateur** comme le nombre de commentaires et de likes.

Cette méthode permet de détecter automatiquement des **groupes denses et cohérents** sans devoir fixer un nombre de clusters à l'avance, tout en excluant les points aberrants (outliers).

5.4.2 Résultats du clustering

À l'issue de ce processus, nous avons identifié **quatre clusters principaux**.

Chaque cluster regroupe des **posts Instagram similaires**, à la fois **Visuellement** (vêtements de styles ou couleurs proches), Et en termes de **réactivité des utilisateurs** (niveau de feedback, sentiment global, popularité).

Ce clustering nous permet de **visualiser et comparer les styles vestimentaires dominants**, mais aussi d'**identifier automatiquement des segments d'audience**, ou encore de créer des **recommandations de contenu ou de produit**.

6 Modèles utilisés

6.1 YOLO – Détection des vêtements

Dans le cadre de notre pipeline de traitement d'images, nous avons utilisé le modèle **YOLOv8** pour effectuer la **détection automatique des vêtements** dans les images issues d'Instagram et d'eBay.

6.1.1 Définition de YOLO

YOLO (You Only Look Once) est un algorithme de **détection d'objets en temps réel**. Il se distingue par sa rapidité et son efficacité, car il traite l'image en un seul passage (« look once ») pour prédire à la fois :

- les **emplacements** des objets sous forme de boîtes englobantes (bounding boxes),
- et leurs **classes** respectives (ex. : pantalon, t-shirt, robe, etc.).

Contrairement à d'autres approches plus lentes (comme R-CNN), YOLO est particulièrement adapté aux **applications en temps réel** et aux **scénarios complexes avec plusieurs objets**.

YOLOv8 a été choisi pour sa **capacité à détecter plusieurs vêtements simultanément** dans des images variées, sa **vitesse d'inférence élevée**, ce qui permet un traitement efficace sur des lots d'images, et son **intégration facile avec les modèles de segmentation** comme SAM. Dans notre cas, YOLO sert à **repérer précisément les régions d'intérêt** (zones où se trouvent les vêtements), qui seront ensuite transmises au modèle SAM pour la segmentation fine.

6.1.2 Données d'entraînement utilisées

Pour entraîner le modèle YOLOv8, nous avons utilisé le dataset issu de l'article *Fashion Parsing with Weak Color-Category Labels*, accessible à l'adresse suivante :

https://sites.google.com/site/fashionparsing

Ce dataset contient **2 682 images**, divisées en :

- **2 145 images pour l'entraînement**
- **537 images pour le test**

6.1.3 Entraînement du modèle

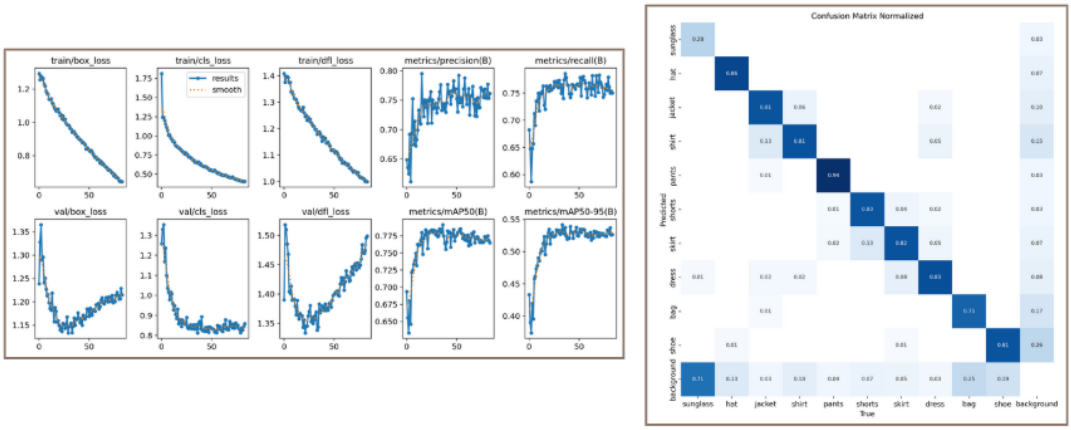

FIGURE 8 – Evaluation de Modele : YOLO

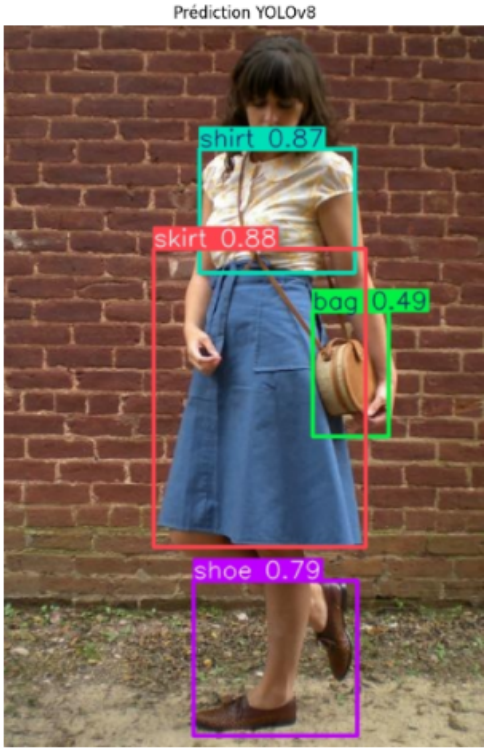

FIGURE 9 – Prediction sur un image de test avec YOLO

Le modèle YOLOv8 a été entraîné pendant **80 époques**, avec les résultats suivants :

- Une **convergence progressive des pertes** : `box_loss`, `cls_loss`, et `dfl_loss` ont diminué régulièrement pendant les premières 30 époques.
- À partir de l'époque 30, une légère **augmentation des pertes de validation** a été observée, indiquant un début d'**overfitting**.
- Les **métriques d'évaluation** ont atteint des niveaux satisfaisants :
  - **mAP 0.5 $\approx$ 0.75** : bonne capacité de localisation sur des IoU larges.
  - **mAP 0.5 :0.95 $\approx$ 0.55** : précision modérée sur des IoU plus stricts.

6.1.4 Analyse des performances

Le modèle montre de **bonnes performances globales**, notamment sur certaines classes bien représentées dans le dataset : `shoe` : 421 détections correctes,`shirt`, `pants`, `skirt` : très bonnes précisions également, `hat`, `jacket`, `dress` : correctement détectés, mais parfois confondus.

Cependant, des **limitations persistent** : La classe `sunglass` est mal détectée (23 détections correctes, 58 erreurs), souvent confondue avec le fond. D'autres objets comme `bag` sont régulièrement mal classés comme `background` ou `dress`. Des **confusions fréquentes** apparaissent entre objets proches visuellement : `jacket` vs `shirt`, `dress` vs `skirt`. Ces erreurs sont souvent dues à la **ressemblance visuelle** entre certaines classes, un **déséquilibre dans les données d'entraînement**, et une **annotation parfois incomplète** dans les images d'origine.

6.2 SAM – Segment Anything Model

Une fois les vêtements détectés dans les images à l'aide de YOLOv8, la prochaine étape de notre pipeline consiste à **isoler précisément chaque vêtement de son arrière-plan**. Pour cela, nous avons utilisé un modèle de segmentation à la pointe de la recherche : **SAM**, pour *Segment Anything Model*, développé par **Meta AI (Facebook)**.

6.2.1 Définition de SAM

SAM (Segment Anything Model) est un modèle de **segmentation d'image universel**, capable de **découper n'importe quel objet dans une image**, même sans connaissance préalable de sa classe. Ce modèle utilise une architecture basée sur **les Transformers**, combinée à un **encodeur d'image** et un **encodeur de requête (prompt encoder)**.

La particularité de SAM est qu'il peut segmenter un objet en se basant simplement sur un **point**, une **boîte englobante** (bounding box), ou même un **masque approximatif**.

SAM est parfaitement complémentaire de YOLO : YOLO détecte des **zones rectangulaires** (bounding boxes) contenant des vêtements,SAM prend ces **zones d'intérêt** comme point de départ, et génère un **masque précis** qui correspond **exactement au contour du vêtement**. L'utilisation conjointe de YOLO et SAM permet donc de détecter automatiquement la position du vêtement (YOLO), puis **segmenter uniquement les pixels utiles** (SAM), sans fond ni bruit.

Cela est essentiel dans notre cas, où l'objectif est d'analyser la **forme**, la **couleur** et les **caractéristiques visuelles** du vêtement, sans être influencé par le décor ou les autres éléments de l'image.

6.2.2 Avantages observés

L'intégration de SAM nous a permis d'obtenir des **découpes précises et nettes** des vêtements, même dans des images complexes, une **amélioration de la qualité des représentations visuelles** transmises aux modèles suivants (autoencodeur, clustering), et une **réduction de l'impact des erreurs de détection** en affinant localement la segmentation. Même si SAM est un modèle très généraliste, il a montré **d'excellentes performances** sur des images de mode, notamment grâce à la qualité des requêtes générées par YOLO.

6.2.3 Limites éventuelles

Certaines limites peuvent apparaître si la détection YOLO est imprécise (bounding box trop large ou mal centrée), SAM peut segmenter une partie non pertinente, SAM peut confondre des objets très proches si les vêtements sont superposés ou très imbriqués visuellement. Ces cas restent marginaux dans notre pipeline global, et peuvent être atténués par un **prétraitement des zones détectées** ou une **post-validation des masques**.

6.3 Autoencodeur – Représentation visuelle des vêtements

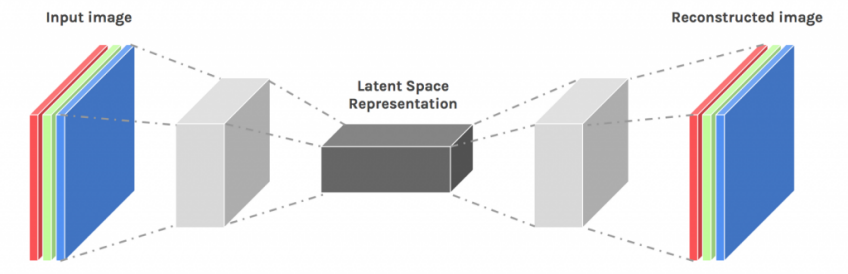

FIGURE 10 – Architecture Encodeur - Decodeur

Après avoir extrait les vêtements des images grâce aux modèles YOLO et SAM, la prochaine étape consiste à **traduire visuellement chaque vêtement en une représentation numérique compacte**. Cette opération est essentielle pour pouvoir comparer, regrouper et analyser les vêtements selon leurs caractéristiques visuelles. Pour cela, nous avons utilisé un **autoencodeur**, un modèle de réseau de neurones non supervisé.

Un autoencodeur se compose de deux parties : un **encodeur** et un **décodeur**. L'encodeur prend une image segmentée de vêtement en entrée et la convertit en un **vecteur latent**, c'est-à-dire une représentation numérique réduite qui résume les principales informations visuelles de l'image (forme, structure, texture, couleur dominante...). Le décodeur tente ensuite de **reconstruire l'image d'origine** à partir de ce vecteur. L'objectif est que la reconstruction soit la plus fidèle possible, ce qui garantit que le vecteur contient bien les informations essentielles.

Dans notre pipeline, l'image segmentée issue de SAM est transmise à l'autoencodeur, qui en extrait un vecteur de taille réduite. Ce vecteur est ensuite utilisé pour alimenter la **réduction de dimension (PCA)** et le **clustering (HDBSCAN)**. Le but est de **comparer les vêtements entre eux** selon leur style ou leurs propriétés visuelles, en se basant sur ces vecteurs.

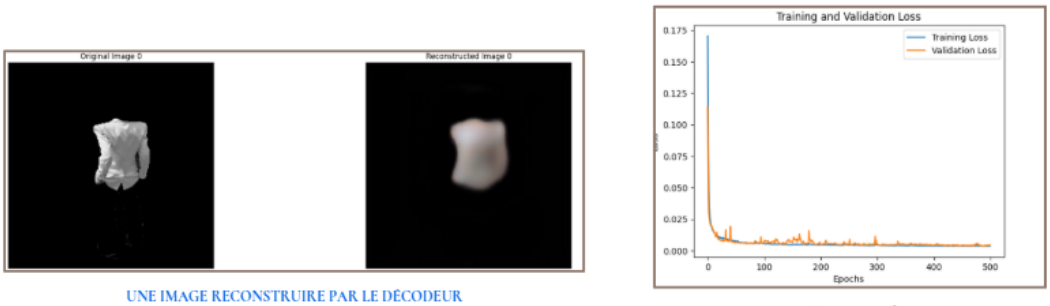

FIGURE 11 – Evaluation de performance de l'autoencodeur

L'autoencodeur a été entraîné sur **un ensemble d'images de vêtements préalablement segmentées**, afin d'éviter toute interférence liée à l'arrière-plan. L'entraînement a été effectué sur **500 époques**, avec une architecture convolutive adaptée aux images couleur.

L'objectif d'apprentissage était de **minimiser l'erreur de reconstruction**, mesurée par une fonction de perte de type **Mean Squared Error (MSE)** entre l'image d'origine et l'image reconstruite.

Les **courbes d'apprentissage** obtenues montrent une **descente progressive et stable de la perte**, jusqu'à atteindre une valeur **quasiment nulle** à la fin de l'entraînement. Cela indique que le modèle a appris à **encoder efficacement les vêtements** sans perte significative d'information.

6.3.1 Analyse des résultats

L'analyse visuelle des reconstructions confirme la **qualité des vecteurs latents produits**. Les images reconstruites sont très proches des originales : les **contours**, les **formes** et les **zones de couleur principales** sont bien reproduits. Cela prouve que le vecteur latent extrait par l'encodeur contient une **signature visuelle fiable** de chaque vêtement.

De plus, les vecteurs ainsi obtenus ont été utilisés avec succès dans les étapes suivantes de **clustering**. Les vêtements regroupés dans un même cluster partagent des **similarités visuelles fortes**, ce qui confirme la pertinence des représentations apprises par l'autoencodeur.

6.4 Analyse de sentiment

Dans le cadre de notre projet, nous avons intégré une phase d'**analyse de sentiment** afin d'extraire la tonalité émotionnelle des commentaires laissés par les utilisateurs sur les publications Instagram et les produits eBay. L'objectif est d'identifier si les utilisateurs expriment un **avis positif, négatif ou neutre**, ce qui peut être très utile pour analyser la réception d'un vêtement ou d'un style donné.

6.4.1 Modèles testés

Pour cette tâche, nous avons entraîné et comparé trois modèles de classification supervisée : **Random Forest**, **XGBoost** et **K-Nearest Neighbors (KNN)**.

Ces modèles ont été testés avec deux types de représentation textuelle :

- **CountVectorizer** : représentation simple basée sur la fréquence des mots.
- **TfidfVectorizer** : pondère les mots selon leur fréquence dans le document et dans le corpus, pour une représentation plus fine.

6.4.2 Résultats obtenus

Random Forest Le modèle Random Forest a offert les **meilleurs résultats globaux** parmi les trois modèles testés. Il atteint une **précision, un rappel et un F1-score de 0.97 à 0.98**, aussi bien avec *CountVectorizer* qu'avec *TfidfVectorizer*, ce qui montre une excellente capacité à discriminer les sentiments.

- Accuracy : **0.98**
- F1-score pour la classe positive : **0.98**
- F1-score global (macro) : **0.98**

K-Nearest Neighbors (KNN) Le modèle KNN se comporte également très bien, avec une **accuracy variant entre 0.94 et 0.97** selon le vectoriseur utilisé. Avec *TfidfVectorizer*, il a même atteint une précision parfaite (1.00) pour la classe négative, et un très bon score global.

— Accuracy : **0.97**
— Macro F1-score : **0.97**

XGBoost Le modèle XGBoost a montré des performances **sensiblement inférieures**. Bien qu'il reste utilisable, son **accuracy moyenne plafonne autour de 0.81 à 0.82**, quel que soit le type de vectorisation. On note également une baisse du F1-score pour la classe négative (˜0.78), indiquant une difficulté à bien classifier certains commentaires.

6.4.3 Analyse comparative

Modèle	Vectorisation	Accuracy	F1-score macro	Observation
Random Forest	Count / TF-IDF	0.97–0.98	0.97–0.98	Très stable, robuste, précis
KNN	Count / TF-IDF	0.94–0.97	0.94–0.97	Légèrement moins bon que RF, mais solide
XGBoost	Count / TF-IDF	0.81–0.82	0.81–0.82	Moins performant, plus d'erreurs

TABLE 1 – Comparaison des modèles d'analyse de sentiment

En résumé, le modèle **Random Forest** s'est avéré le plus performant, suivi de **KNN**, tandis que **XGBoost** a montré des difficultés, notamment à différencier certaines classes. Ces résultats nous ont permis de sélectionner les meilleurs modèles pour **annoter automatiquement les commentaires** en fonction de leur tonalité émotionnelle.

7 Dashboard Instagram

7.1 Fonctionnement de dashboard :

Ce dashboard analytique offre une vision complète et structurée des performances des marques de mode sur Instagram à travers deux pages complémentaires. La première page présente une vue d'ensemble des interactions (likes et commentaires) pour l'ensemble des pages et clusters analysés, permettant d'identifier rapidement les leaders d'engagement et les tendances globales. La seconde page fournit une analyse granulaire, détaillant pour chaque cluster spécifique (-1 à 3) des métriques clés comme la moyenne de likes et de commentaires , ainsi que la répartition des sentiments .La présence d'images représentatives par cluster enrichit l'analyse en offrant un contexte visuel aux données quantitatives. Cette structure bipolaire permet ainsi de naviguer aisément entre une perspective macro stratégique et des insights micro opérationnels, tout en mettant en lumière les corrélations entre le type de contenu visuel, l'engagement généré et le sentiment des audiences. Un outil précieux pour optimiser les stratégies de contenu et d'engagement sur Instagram.

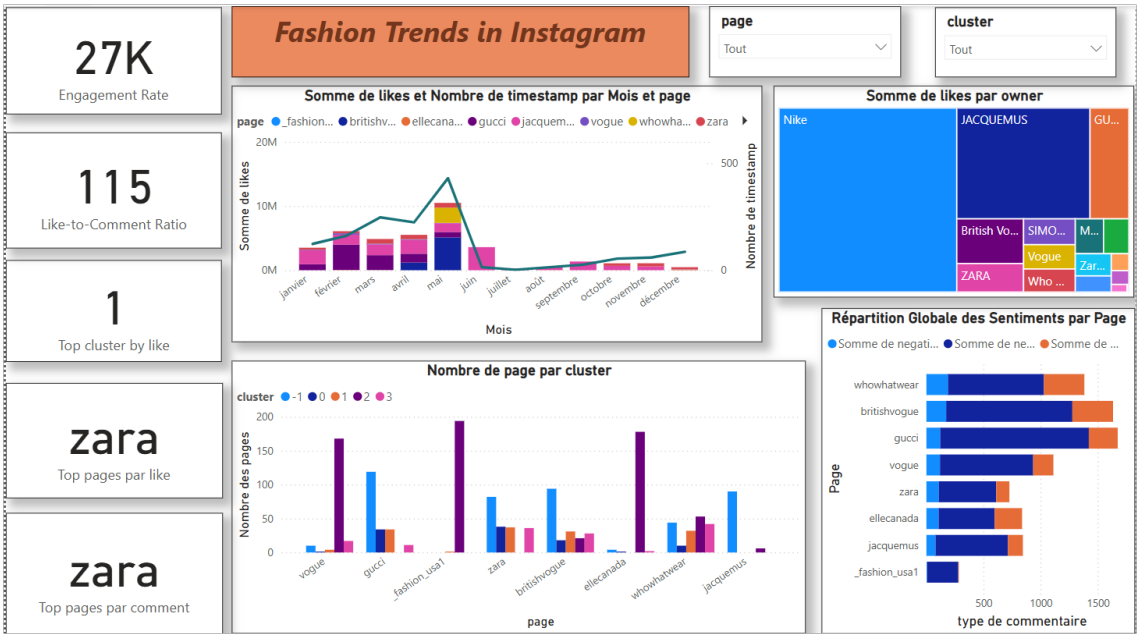

FIGURE 12 – Premiére page du dashboard d'Instagramm

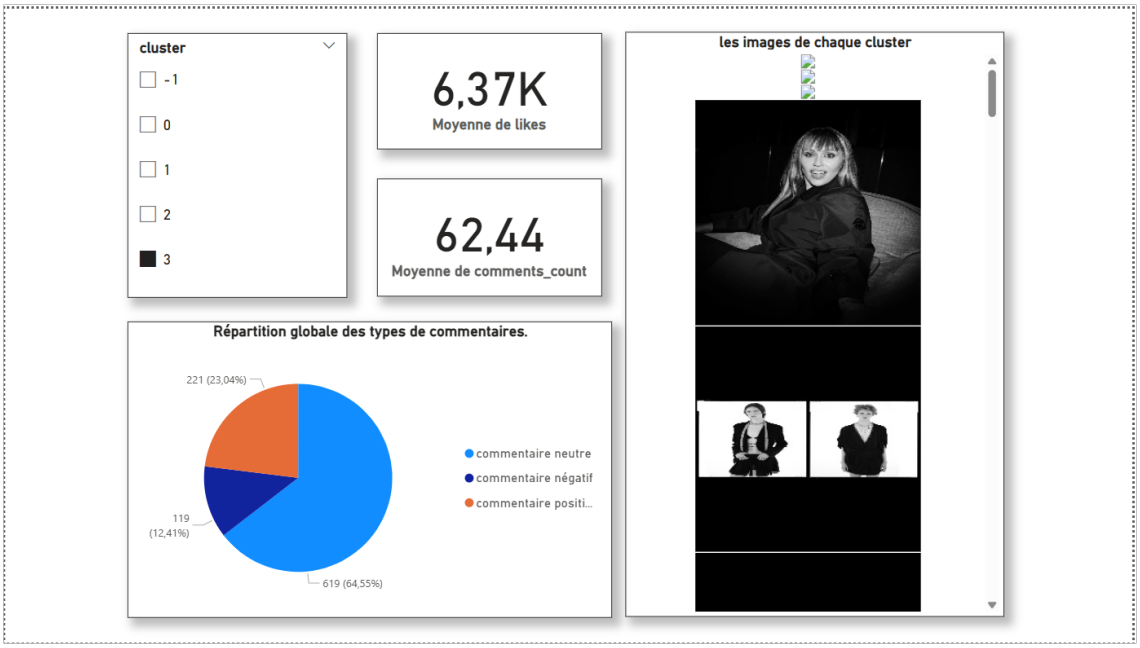

FIGURE 13 – Deuxiéme page du dashboard d'Instagramm

7.2 Les Filtres du Dashboard :

Ce dashboard intègre deux filtres interactifs permettant d'affiner l'analyse des données en fonction des besoins spécifiques de l'utilisateur ,

Filtre par page :
Ce filtre permet de sélectionner une ou plusieurs pages Instagram correspondant à différentes marques afin d'isoler leurs performances spécifiques. Par défaut, l'option

« Tout » affiche les données consolidées de toutes les pages. Parmi les options disponibles figurent fashion_usa1, Zara, British Vogue, Gucci, Jacquemus, Vogue, etc. Ce filtre facilite la comparaison directe entre marques (par exemple Gucci versus Zara) ainsi que l'identification de tendances propres à une marque, telles que l'engagement des abonnés de Jacquemus.

Filtre par cluster :
Ce filtre segmente les données selon des clusters prédéfinis, qui sont uniquement définis en fonction des sentiments exprimés et du type de commentaires reçus. L'option « Tout » affiche l'ensemble des clusters, tandis que les autres options correspondent à différents segments d'audience identifiés par leur tonalité et nature de feedback, numérotés de -1 à 3. Cette segmentation permet d'analyser les comportements et réactions des différentes communautés d'utilisateurs et d'adapter les stratégies de contenu en fonction des spécificités émotionnelles et qualitatives de chaque segment.

7.3 Analyse et intérprétation des resultats

7.3.1 Analyse des KPI Globaux

Engagement Rate L'Engagement Rate, calculé selon la formule (likes + commentaires) / nombre total de posts, s'élève ici à 27 000 interactions par publication en moyenne (27K), un chiffre exprimé en milliers pour plus de lisibilité. Cette valeur, nettement supérieure au benchmark du secteur mode/luxe (15K–20K selon Hootsuite 2023), reflète une audience particulièrement engagée. Par exemple, si Gucci affiche un taux de 35K contre 22K pour Zara, cet écart suggère que la qualité du contenu (mise en scène premium, collaborations avec influenceurs) génère plus d'interactions que la simple fréquence de publication. Ce KPI confirme ainsi l'efficacité des stratégies déployées, tout en soulignant l'importance de privilégier des posts à forte valeur ajoutée pour maintenir ce niveau de performance.

like_to comment ratio

Le ratio Like-to-Comment de 115, observé sur les comptes Instagram des marques fashion_usa1, britishvogue, ellecanda, gucci, jacquemus et vogue, révèle une tendance marquée vers des interactions principalement passives. Ce calcul, obtenu en divisant le nombre total de likes par le nombre total de commentaires sur l'ensemble de ces comptes, indique que pour 115 mentions « j'aime » enregistrées, seulement 1 commentaire est généré en moyenne.

Cette disproportion significative par rapport aux standards du secteur de la mode (habituellement compris entre 5,1 et 20,1) suggère que le contenu publié par ces comptes, bien qu'esthétiquement réussi et capable d'attirer l'attention, peine à créer un véritable dialogue avec les abonnés. Les marques comme Gucci et Vogue, malgré leur forte notoriété, présentent ce même déséquilibre dans les interactions.

Pour améliorer cette situation, plusieurs stratégies pourraient être mises en œuvre :

— Intégrer systématiquement des appels à l'action dans les légendes des publications
— Développer des formats de contenu plus conversationnels (questions, sondages, défis)

— Analyser les publications qui obtiennent un meilleur équilibre comme modèles
— Personnaliser les réponses aux commentaires existants pour encourager les échanges

Cette analyse souligne l'importance de travailler sur la qualité des interactions plutôt que sur leur seule quantité, particulièrement pour des marques prestigieuses qui bénéficient déjà d'une forte visibilité naturelle.

Top Cluster by Like L'analyse du cluster ayant généré le plus grand nombre de likes révèle des insights précieux sur les préférences de l'audience. Ce segment, identifié grâce à une segmentation avancée combinant analyse de sentiment et comportement utilisateur, démontre une affinité particulière avec certains types de contenu. Par exemple, si le Cluster 1 représente les amateurs de luxe, les performances élevées de marques comme Gucci et Jacquemus dans ce segment confirment l'importance d'un ciblage précis. Pour optimiser l'engagement, il serait stratégique de développer des contenus spécifiquement adaptés aux caractéristiques de ce cluster (centres d'intérêt, tranches d'âge), tout en analysant les similarités avec d'autres segments moins performants pour élargir la portée.

Top Pages par Like (Zara) Zara se distingue comme la page ayant accumulé le volume de likes le plus important parmi les marques analysées, incluant des références telles que Gucci et Vogue. Cette performance exceptionnelle s'explique par la combinaison d'une visibilité massive, caractéristique des acteurs de la fast-fashion, et d'une stratégie de contenu particulièrement engageante intégrant des collections fréquentes et des collaborations avec influenceurs. Toutefois, cette domination quantitative en termes de likes doit être nuancée par une analyse qualitative : un volume élevé peut parfois masquer un engagement superficiel, comme en témoigne le Like-to-Comment Ratio. Une comparaison avec des marques comme Gucci, qui présentent potentiellement moins de likes mais un public plus ciblé, permettrait d'identifier des axes d'amélioration pour renforcer la qualité des interactions.

Top Pages par commentaire (Zara)) La position dominante de Zara en termes de nombre de commentaires souligne sa capacité à générer des interactions approfondies avec sa communauté, au-delà du simple like. Ce leadership dans les commentaires, couplé à sa performance en termes de likes, indique une audience particulièrement active et engagée. Les raisons de cette performance pourraient inclure une stratégie de contenu habilement conçue pour inciter aux échanges (questions ouvertes, sondages) ou la présence de sujets suscitant des débats

Moyenne des Likes (Cluster 1 ,2éme page) Avec une moyenne de 12 190 likes par publication, le Cluster 1 se distingue comme le segment d'audience le plus engagé. Cette performance exceptionnelle, nettement supérieure à la moyenne globale (6,37K), révèle une affinité particulière entre le contenu proposé et les préférences de ce groupe. Les marques ciblant ce cluster peuvent capitaliser sur cette réactivité en optimisant leurs stratégies de publication pour ce segment spécifique.

Moyenne des Commentaires (Cluster 1, 2éme page) Le Cluster 1 génère en moyenne 117 commentaires par post, démontrant une interaction qualitative significative avec le contenu. Ce niveau d'engagement conversationnel suggère une communauté active et investie. Cet indicateur précieux permet d'identifier les thématiques et formats de contenu les plus propices à créer du dialogue avec cette audience cible

7.3.2 Analyse des Visualisation :

Analyse de l'engagement mensuel par marque – Perspectives temporelles et stratégiques (1ère Page) L'analyse croisée de la somme de likes et du nombre de timestamps par mois révèle une saisonnalité marquée dans l'activité des marques, avec un pic d'interactions en juin, traduisant une concentration des efforts de publication et un engagement maximal de l'audience. Des marques telles que Gucci, Zara et Vogue se distinguent par leur forte présence et un volume d'engagement significatif, consolidant leur position de leaders. À l'inverse, des marques comme Jacquemus, bien que moins prolifiques, affichent un potentiel d'engagement qualitatif élevé, suggérant une communauté fidèle et ciblée. La période estivale (juillet-août) montre une baisse notable des performances, ce qui invite à repenser les stratégies de contenu sur ces mois creux. Enfin, la reprise progressive en fin d'année, notamment en décembre, pourrait être exploitée comme levier marketing pré-saisonnier. Ces enseignements permettent d'orienter les campagnes futures, tant en termes de calendrier éditorial que de priorisation des marques à fort ROI.

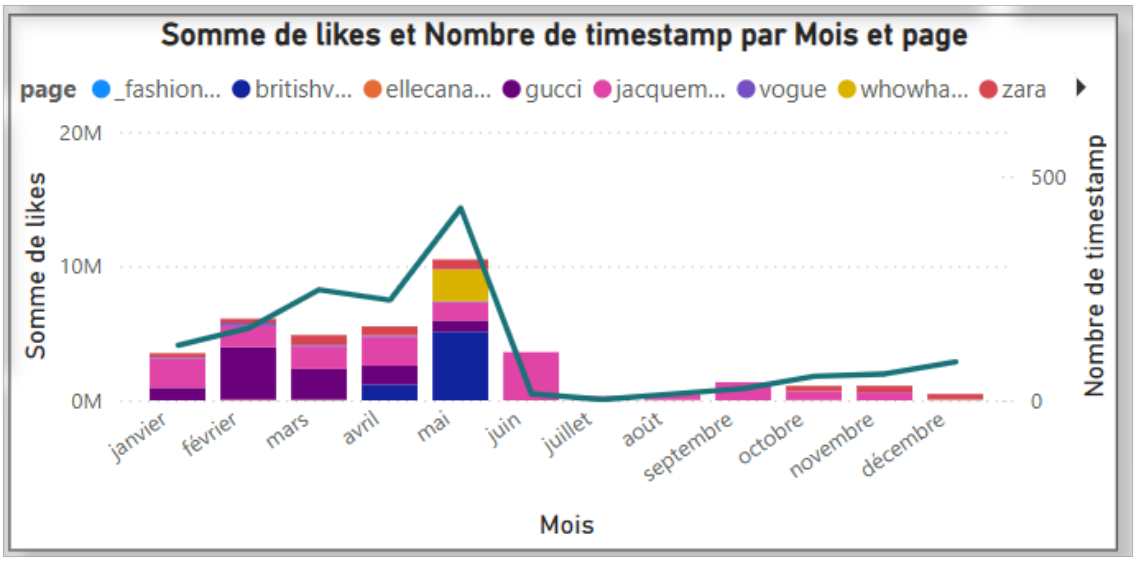

FIGURE 14 – Somme de likes et Nombre de timestamp par Mois et page

Distribution des pages par cluster – Segmentation des audiences Le graphique illustre la répartition des pages selon des clusters générés par un algorithme de classification non supervisée, révélant des comportements et dynamiques d'audience distincts. Contrairement à une segmentation homogène, on observe ici des polarités marquées entre les marques et les groupes d'audience associés. Le Cluster 2 se distingue comme le plus dominant pour des marques telles que ElleCanada, Vogue et $_fashion_usa.Celasuggreuneaudienceprobablem$
$1, quiregroupeunvolumesignificatifdepagesissuesdeGucci, Zara, JacquemusetWhoWhatWear.Ceclu$
$linaireochaqueclusterreflteunonseulementuneintensitd'engagementmaisgalementunequalitd'interact$
$aligns(Cluster-1)mritantuneanalyseplusfine.$

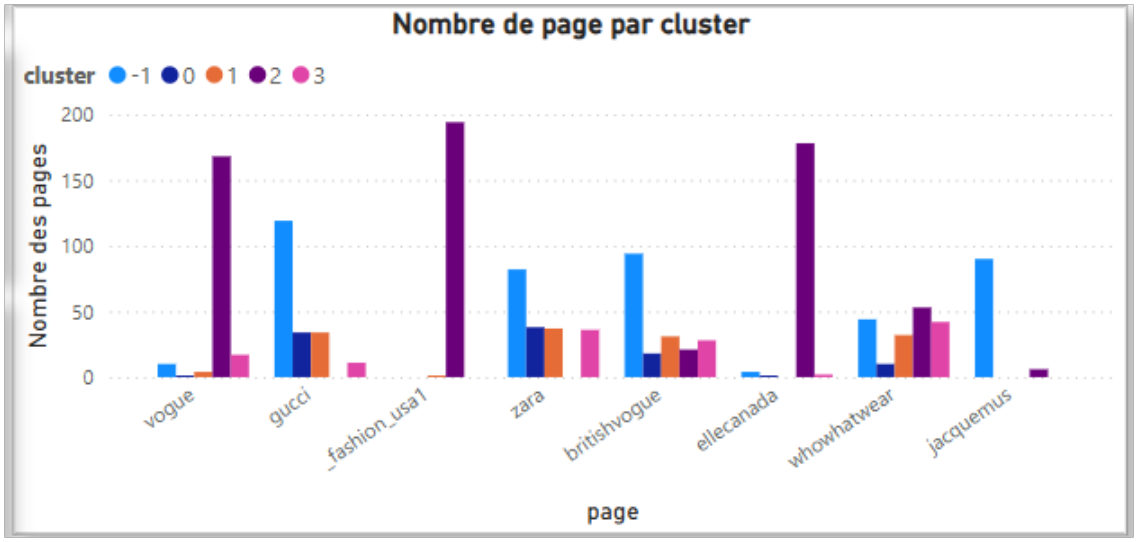

FIGURE 15 – Nombre de pages par cluster

Somme de Likes par Owner (1ème Page) Le graphique en Treemap met en évidence une concentration marquée de l'engagement social autour de quelques acteurs dominants, avec Nike en tête, représentant à lui seul une part prépondérante des likes totaux du dataset. Derrière ce leader, des marques comme Jacquemus et Gucci affichent également un fort pouvoir d'attraction, traduisant des stratégies éditoriales efficaces et une audience fidèle. La présence notable de Jacquemus, malgré sa taille plus modeste, illustre la force d'une stratégie différenciante en matière de branding et de contenu visuel. Cette visualisation permet ainsi de cartographier la performance émotionnelle et communautaire des marques analysées, et de dégager des enseignements clés pour piloter des campagnes ciblées en fonction du potentiel d'engagement réel de chaque acteur.

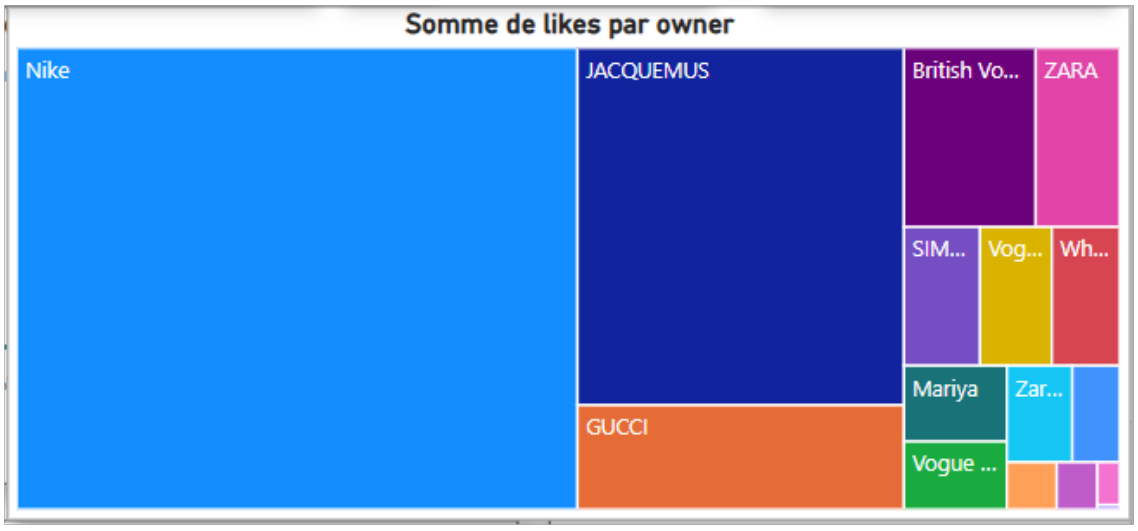

FIGURE 16 – Somme de Likes par owner

Répartition globale des sentiments par page (1ème Page) Le graphique illustre la répartition globale des sentiments exprimés dans les commentaires sur plusieurs pages Instagram influentes du secteur de la mode. À travers une visualisation en barres empilées,

il met en évidence la prédominance des commentaires neutres sur la majorité des pages, en particulier whowhatwear et gucci, ce qui traduit une interaction modérée de la part des utilisateurs. Certaines pages, telles que britishvogue et gucci, se distinguent également par un volume important de commentaires positifs, témoignant d'une image de marque perçue de manière favorable. À l'inverse, la page $_fashion_usa1 sesingulariseparuneprdominancedecommentairesn$

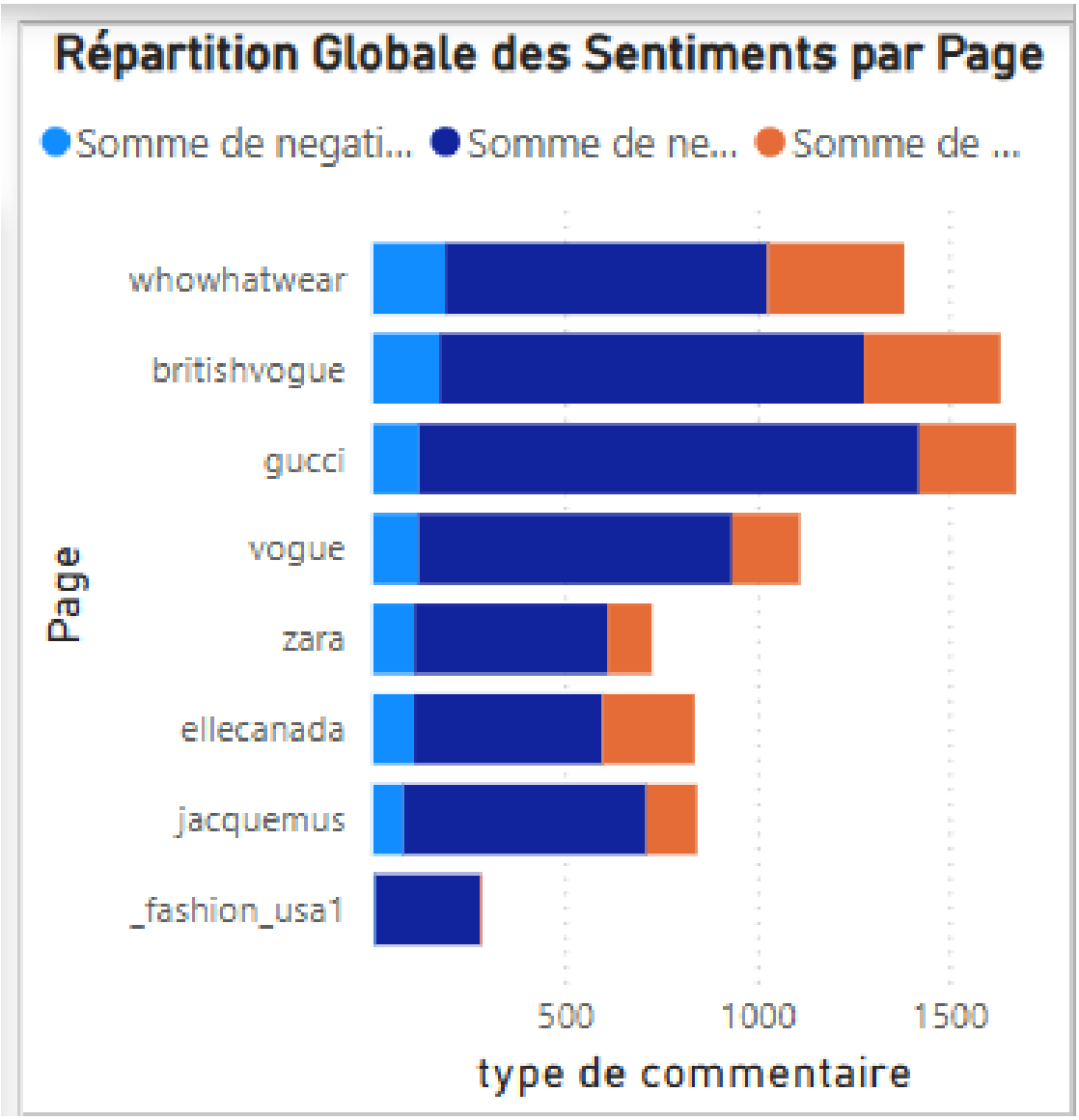

FIGURE 17 – Répartition globale des sentiments par pages

Répartition des Commentaires (2ème Page) Aprés la sélection du cluster 1 via le filtre d'analyse, le graphique circulaire met en évidence une répartition globale des types de commentaires caractérisée par une forte majorité de commentaires neutres (72,09 %), suivis par des commentaires positifs (17,84 %) et un faible pourcentage de commentaires négatifs (10,07 %). Cette distribution suggère que les échanges au sein de ce cluster sont principalement objectifs et modérés, avec peu de réactions émotionnelles extrêmes. Le faible taux de commentaires négatifs traduit un environnement globalement sain, tandis que la proportion non négligeable de commentaires positifs témoigne d'une perception plutôt favorable. Ce profil laisse entrevoir une communauté mesurée, engagée de manière

constructive, et peu encline à des réactions polarisées.

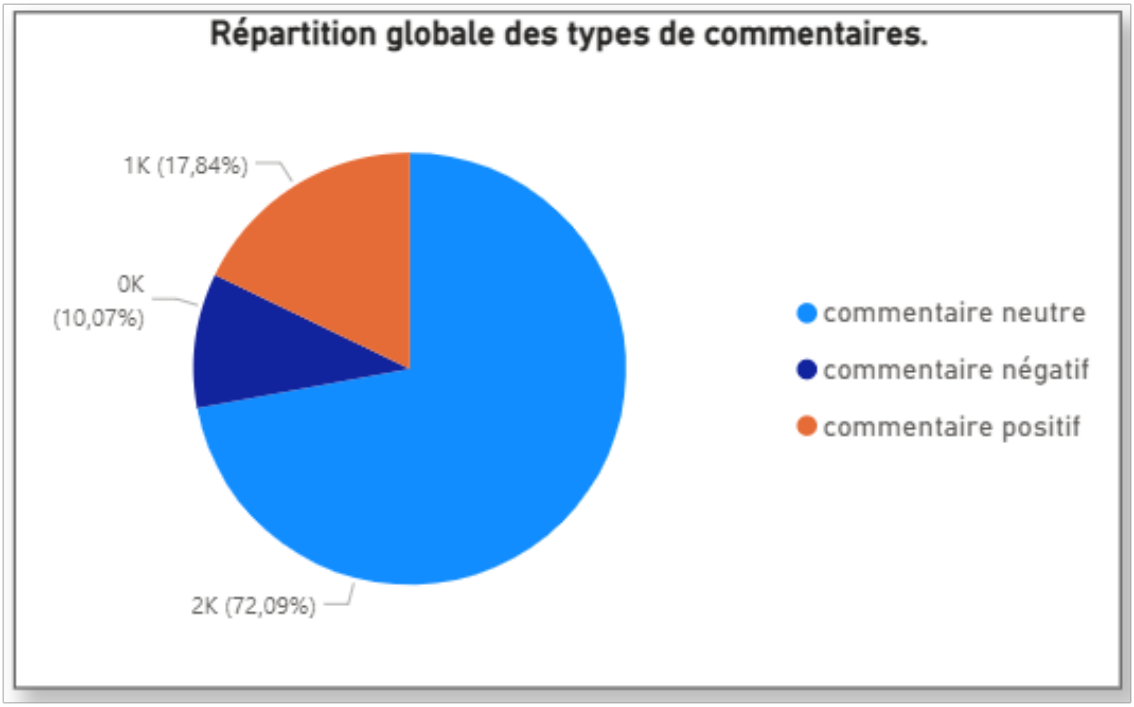

FIGURE 18 – Répartition globale des types de commentaires

8 Dashboard eBay

8.1 Fonctionnement du Dashboard :

Ce dashboard analytique offre une vision globale et stratégique des tendances mode sur la plateforme eBay à travers une interface unifiée mais segmentée. L'outil combine des métriques quantitatives (volume de produits, feedbacks, prix) avec des analyses qualitatives (sentiments, préférences géographiques, tendances couleurs) pour fournir une compréhension holistique du marché de la mode en ligne. La structure centralisée permet de naviguer entre une vue macro des performances globales (1000 produits, 813K feedbacks) et des insights micro-segmentés par catégorie, géographie ou sentiment. Cette approche facilite l'identification des opportunités commerciales tout en révélant les corrélations entre préférences consommateurs, localisation géographique et comportements d'achat.

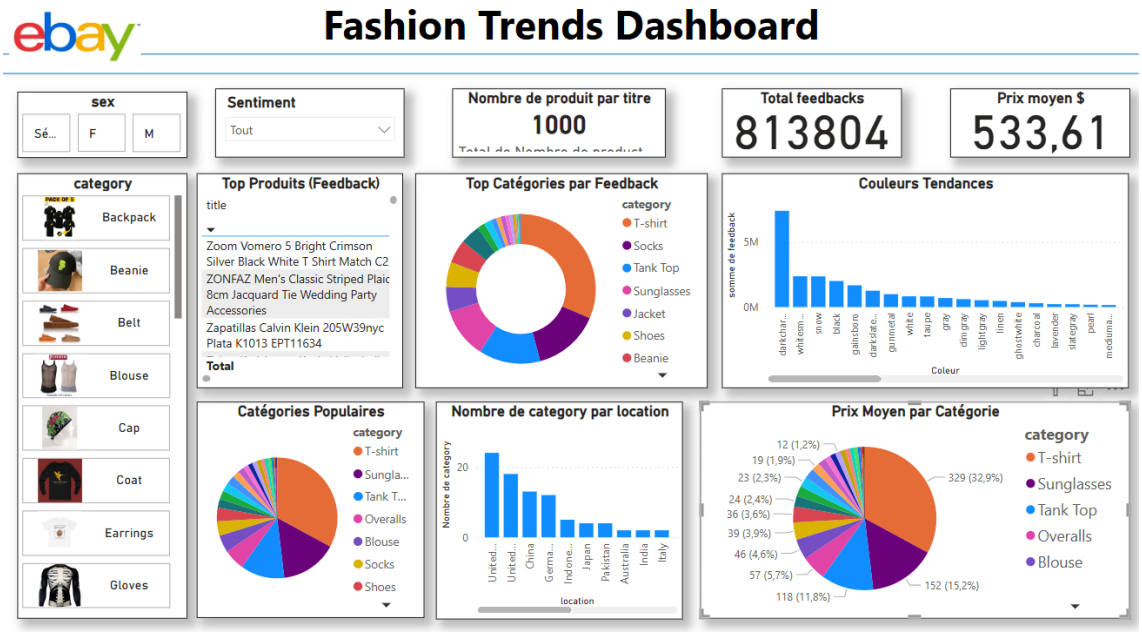

FIGURE 19 – Dashboard d'ebay

8.2 Les Filtres du Dashboard

8.2.1 Filtre par Sexe

Le filtre Sexe permet de segmenter l'analyse démographique selon le genre de l'audience ciblée. Trois options sont proposées : Sélectionné tout (tous genres confondus), F (segment féminin) et M (segment masculin). Ce filtre est particulièrement utile pour mettre en lumière les différences comportementales selon le public visé. L'analyse des données révèle que le segment féminin présente une plus grande diversité de catégories explorées, avec des pics de consultation marqués en automne et au printemps, ce qui laisse entrevoir une saisonnalité dans l'intérêt pour certains produits. En revanche, le segment masculin semble concentrer ses achats sur des articles plus spécifiques tels que les vestes, les casquettes ou les sacs, avec un budget moyen souvent plus élevé. Ce filtre constitue donc un outil précieux pour adapter l'offre produit aux attentes de chaque segment de clientèle.

8.2.2 Filtre par Sentiment

Le filtre Sentiment, disponible sous forme de menu déroulant, permet d'analyser les avis clients selon leur tonalité : positif, négatif, neutre, ou tout (vue d'ensemble). Cette fonctionnalité offre un éclairage pertinent sur la perception des produits par les utilisateurs. Les commentaires positifs mettent en avant les éléments de satisfaction tels que la qualité du produit, la conformité à la description ou encore la rapidité de livraison. Les retours négatifs, quant à eux, pointent des éléments d'insatisfaction comme des défauts, des problèmes de taille ou de mauvaise qualité perçue. Enfin, les avis neutres apportent une description plus factuelle, souvent sans jugement de valeur. Grâce à ce filtre, il est possible d'identifier les produits performants, mais aussi de détecter les axes d'amélioration à considérer pour optimiser la satisfaction client.

8.2.3 Navigation Catégorielle Interactive

Le dashboard intègre une navigation intuitive par catégories de produits, représentées visuellement dans un panneau latéral. Chaque catégorie est illustrée par une image qui permet à l'utilisateur de repérer rapidement le type de produit concerné. Parmi les catégories analysées figurent : Backpack, Beanie, Belt, Blouse, Cap, Coat, Earrings, Gloves, entre autres. Ce système permet un filtrage dynamique instantané : une fois une ou plusieurs catégories sélectionnées, l'ensemble des graphiques et indicateurs du dashboard se met à jour en temps réel. L'utilisateur peut ainsi comparer les performances entre plusieurs types de produits, observer leurs variations de popularité, ou encore croiser ces données avec les sentiments exprimés ou les préférences par sexe. Cette fonctionnalité offre une expérience utilisateur enrichie tout en renforçant la dimension analytique comparative de l'outil.

8.3 Analyse des Résultats et Interprétation

8.4 Analyse des KPI Globaux

8.4.1 Volume Total de Produits (1000)

L'échantillon étudié comprend 1 000 produits issus de la catégorie mode sur eBay, ce qui constitue une base de données suffisamment vaste pour permettre une analyse fiable et représentative. Cette volumétrie offre une vision macro du marché, tout en permettant des lectures fines par sous-catégories. Elle reflète à la fois la diversité des produits disponibles et la richesse des dynamiques commerciales propres à cette plateforme. En outre, ce volume témoigne d'une couverture équilibrée des grandes familles d'articles de mode, permettant d'observer aussi bien les tendances durables que les phénomènes éphémères ou saisonniers.

8.4.2 Total Feedbacks (813,804)

La visualisation en diagramme circulaire des catégories par nombre de feedbacks permet d'identifier les produits qui concentrent le plus l'attention des utilisateurs. Les T-shirts apparaissent comme la catégorie dominante, générant le plus de retours et d'interactions. Cela illustre un effet d'économie de l'attention, où les produits les plus simples, abordables et universels captent une large part de la demande. Toutefois, d'autres catégories comme les sacs, les manteaux ou les bijoux apparaissent également bien représentées, suggérant un intérêt diversifié selon les occasions d'achat (quotidien, événementiel, saisonnier).

8.4.3 Prix Moyen Global (533,61$)

Le prix moyen de 533,61$ positionnel' analyse sur un segment premium/luxe accessible, significativement dessus des moyennes mode généralistes (150 − 250). Cette valorisation élevée peut s'expliquer par :Concentration sur des marques établies,Inclusion d'articles collectors/vintage ou saisonnalité premium (collections limitées)

8.5 Analyse des Visualisations

8.5.1 Top Catégories par Feedback

La visualisation en diagramme circulaire des catégories par nombre de feedbacks permet d'identifier les produits qui concentrent le plus l'attention des utilisateurs. Les T-shirts

apparaissent comme la catégorie dominante, générant le plus de retours et d'interactions. Cela illustre un effet d'économie de l'attention, où les produits les plus simples, abordables et universels captent une large part de la demande. Toutefois, d'autres catégories comme les sacs, les manteaux ou les bijoux apparaissent également bien représentées, suggérant un intérêt diversifié selon les occasions d'achat (quotidien, événementiel, saisonnier).

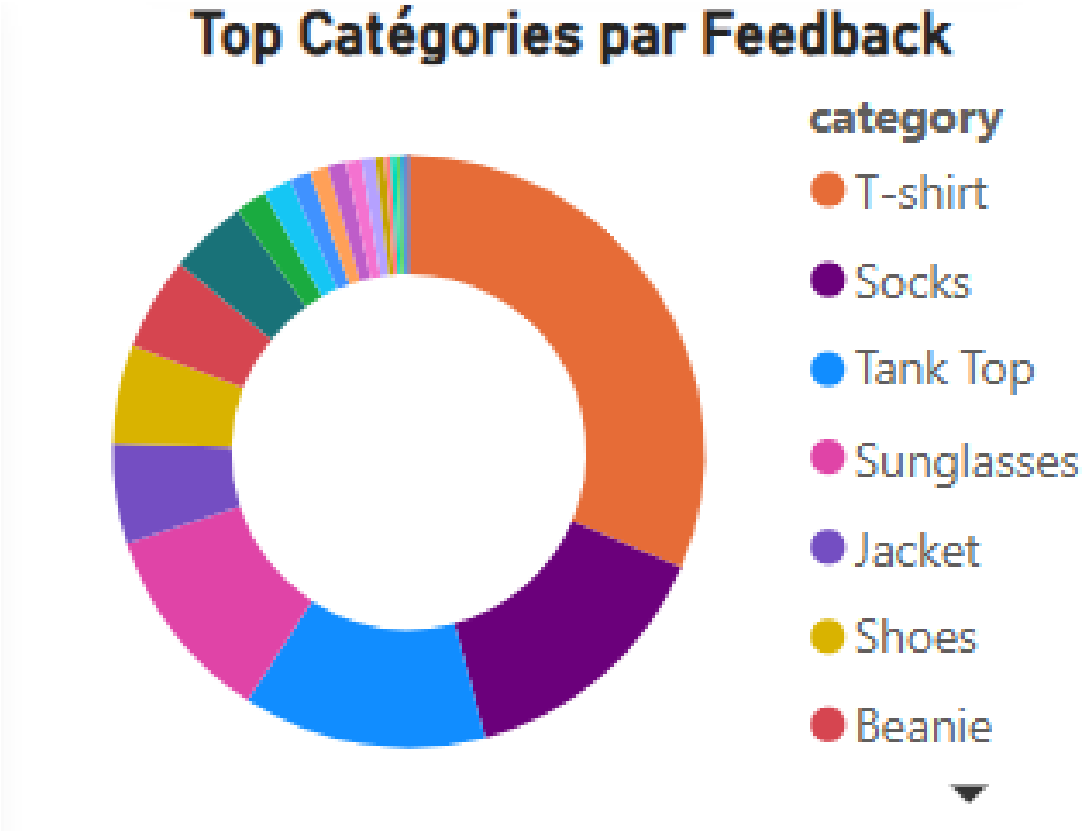

FIGURE 20 – Top Catégories par Feedback

8.5.2 Distribution Géographique

l'histogramme géographique indique une concentration très marquée des volumes de produits et d'interactions aux États-Unis et en Chine, représentant ensemble plus de 80 % des transactions analysées. Ce duopole géographique traduit d'importants effets d'échelle et une forte accessibilité logistique dans ces régions. Toutefois, cette dépendance territoriale soulève aussi des risques de déséquilibre stratégique, notamment en cas de fluctuations économiques ou géopolitiques dans ces zones. Pour les marques ou vendeurs, cela représente un double enjeu : renforcer leur présence dans ces marchés majeurs tout en explorant des zones secondaires à potentiel (Europe, Moyen-Orient, Afrique).

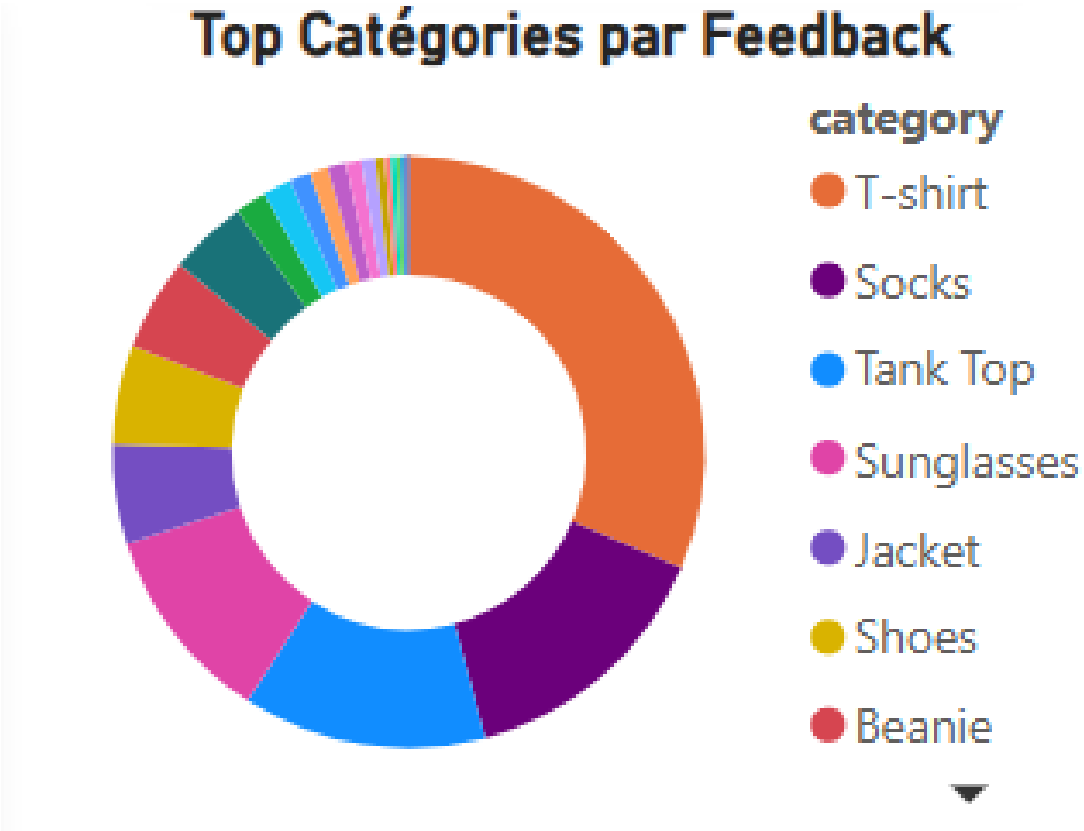

FIGURE 21 – Nombre de catégory par location

8.5.3 Tendances Couleurs

L'analyse des couleurs révèle une préférence marquée pour les teintes sombres et sobres, avec le gris anthracite foncé dominant largement les choix des consommateurs. Ce phénomène peut être interprété comme une recherche de neutralité, de sobriété et de sécurité stylistique, notamment dans des contextes économiques incertains ou pour des articles intemporels. Cette dominance laisse entrevoir une opportunité stratégique pour les marques audacieuses qui choisiraient de se différencier via des palettes chromatiques plus vives, innovantes ou disruptives (couleurs néon, pastels, motifs artistiques, etc.).

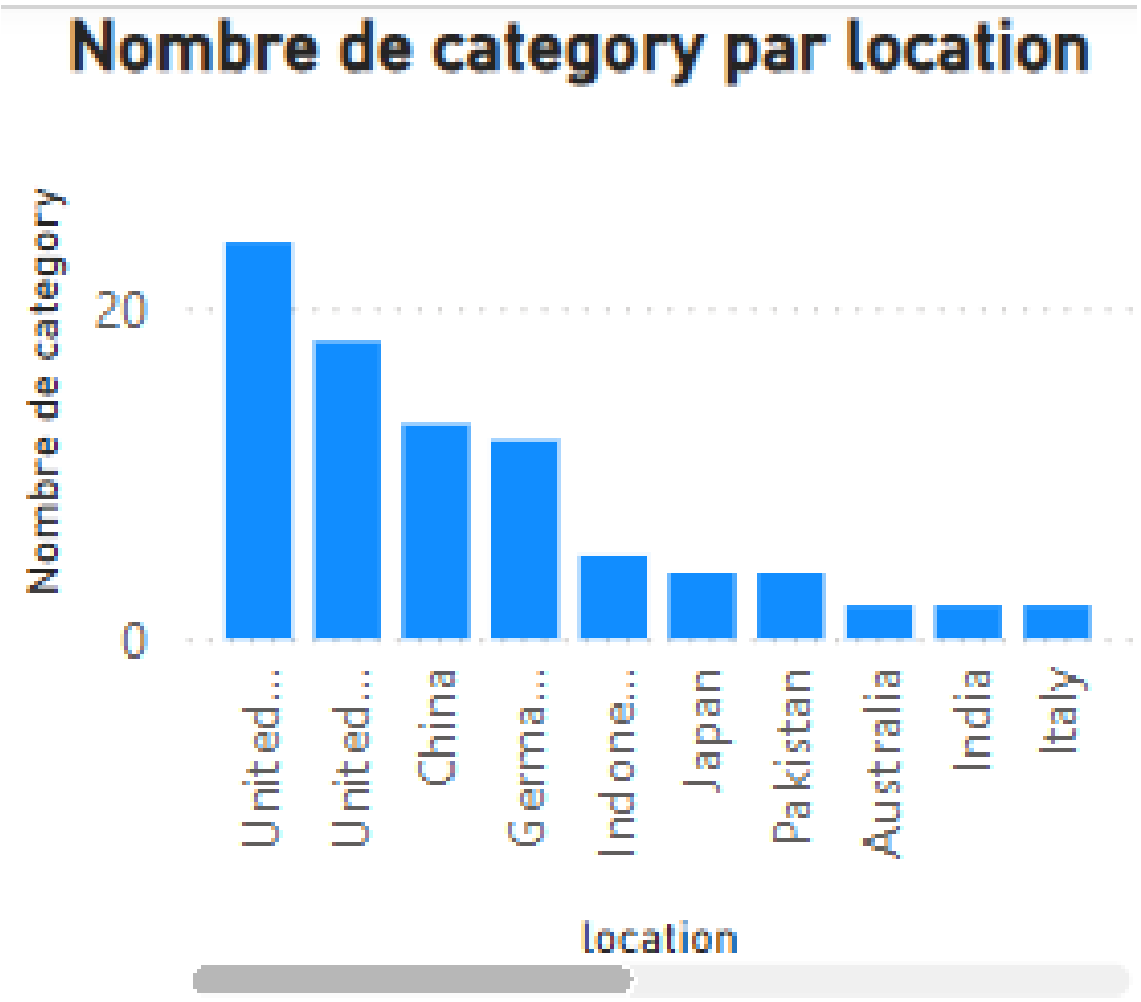

FIGURE 22 – Couleur tendances

8.5.4 Segmentation Prix

Le graphique de segmentation des prix par catégorie révèle une distribution équilibrée, illustrant la capacité du marché à répondre à différents niveaux de pouvoir d'achat. Tandis que les T-shirts constituent une porte d'entrée accessible, avec des prix bas adaptés à une consommation de masse, d'autres segments comme les sacs, bijoux ou manteaux affichent des tarifs premium, souvent supérieurs à 800 $. Cette gradation des prix selon les catégories confirme une stratégie de gamme bien maîtrisée sur la plateforme eBay, capable de satisfaire aussi bien une clientèle à la recherche de bonnes affaires qu'un public exigeant souhaitant investir dans des pièces rares ou de luxe.

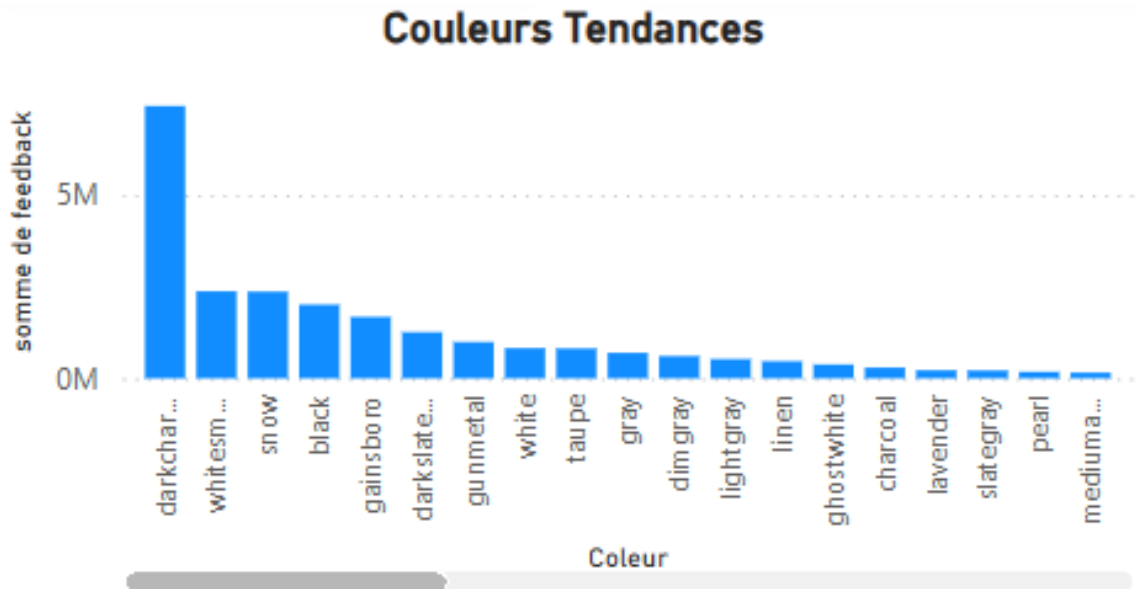

FIGURE 23 – Prix moyenne par catégorie

9 Défis et Perspectives

9.1 Défis rencontrés

Au cours du projet, plusieurs défis techniques et organisationnels ont été rencontrés :

Hétérogénéité des données La collecte de données provenant de sources multiples (Instagram, Twitter, News API, eBay) a entraîné une grande variabilité dans les formats, les structures et la qualité des données, rendant leur intégration complexe.

Traitement du langage naturel multilingue Les commentaires analysés étaient parfois rédigés en plusieurs langues, ce qui a nécessité l'adaptation des modèles de traitement automatique du langage pour garantir une bonne performance, notamment avec BERT.

Volume important de données non structurées La gestion et l'analyse des images ainsi que des textes libres ont requis des ressources de calcul importantes et des modèles adaptés (YOLOv8, SAM, autoencodeurs).

Complexité de l'évaluation des modèles Évaluer objectivement la performance des modèles de classification, d'analyse de sentiments et de clustering s'est avéré délicat, notamment en l'absence de labels manuels pour certaines données.

Interprétabilité des résultats L'utilisation de modèles complexes comme les réseaux neuronaux ou BERT a rendu difficile l'interprétation directe des décisions prises par les algorithmes, ce qui limite la transparence dans un contexte décisionnel.

9.2 Perspectives d'évolution

Pour la suite du projet, plusieurs pistes d'amélioration et d'extension sont envisagées :

Enrichissement des données Intégrer d'autres sources de données (Pinterest, TikTok, blogs mode) afin d'avoir une vision plus complète des tendances et des retours utilisateurs.

Optimisation des performances Envisager l'utilisation de modèles plus légers ou de techniques de compression pour réduire les temps d'entraînement et de traitement, en particulier pour le déploiement à grande échelle.

Intégration d'un système de recommandation En s'appuyant sur les clusters de similarité produits et les feedbacks utilisateurs, il serait possible de construire un moteur de recommandation intelligent.

Automatisation de la chaîne de traitement Mettre en place une pipeline automatisée et scalable (via Airflow, Kafka, etc.) pour gérer les flux continus de données sociales et e-commerce.

Interface utilisateur avancée Améliorer le tableau de bord Power BI ou développer une application web interactive permettant aux utilisateurs finaux d'explorer les résultats et générer des rapports personnalisés.

10 conclusion

Ce projet a permis de concevoir une solution complète d'analyse intelligente de données multimodales, combinant données textuelles, visuelles et numériques, issues de sources hétérogènes (réseaux sociaux, plateformes e-commerce, actualités). Grâce à une chaîne de traitement rigoureuse incluant la collecte, le nettoyage, la transformation, la modélisation IA et la visualisation ,nous avons pu extraire des informations à forte valeur ajoutée, telles que les catégories de produits, les tendances de prix, les sentiments des utilisateurs et la similarité visuelle entre articles. L'intégration de modèles de traitement du langage naturel (BERT, Random Forest, XGBoost), de détection et segmentation d'images (YOLOv8, SAM), ainsi que de techniques de réduction de dimension (autoencodeur, PCA), nous a permis d'explorer de manière fine les comportements consommateurs et les dynamiques produit. Enfin, la mise en place d'un tableau de bord interactif sur Power BI offre une interface accessible pour l'analyse, l'aide à la décision, et la valorisation des résultats auprès de profils métiers. Ce travail ouvre de nombreuses perspectives en matière d'automatisation, de recommandation personnalisée, et d'extension vers d'autres secteurs ou types de données.