Groupe 7 : AIT RADI Ouiza , ZADDI Taous, Cheick Tidiane Sissoko , Thelma LUM, Mai DAO, Jude KALENGAYI

Analyse des Avis sur une chaussure de running sur Decathlon

1. Présentation du sujet :

Nous avons choisi d'analyser les commentaires en ligne concernant les chaussures de running sur decathlon. L'objectif est de déterminer si les avis écrits correspondent aux notes attribuées et d'identifier les tendances générales dans les retours des clients (par exemple la qualité ou à la marque).

Analyse des Avis des chaussures de running sur Decathlon

2. Objectifs :

- Comprendre les principaux facteurs qui influencent la satisfaction des clients (taille etc.).
- Identifier les éventuelles incohérences entre les commentaires et les notes.
- Créer un dashboard pour résumer efficacement les avis des clients.

3. Données nécessaires :

- **Notes attribuées** : sur 5.
- **Commentaire**
- **Date du commentaire** : Pour évaluer d'éventuelles tendances dans le temps.
- **Auteur**

Source :

- Decathlon (section chaussures running).

4. Fonctionnalités prévues :

- **Analyse de sentiment** : Utilisation de modèles d'apprentissage automatique pour déterminer si un commentaire est positif, neutre ou négatif.
- **Classification des évaluations** : Analyser le produit en fonction de leur taux de satisfaction.
- **Tableau de bord** : Visualisation des tendances (exemple : satisfaction des clients sur le produit).

5. Etapes du projet :

1. **Scraping des données** : Collecte des avis en ligne via des outils comme BeautifulSoup et Selenium.
2. **Nettoyage des données** :
    - Normalisation de la colonne "Auteur"
    - Conversion des dates
    - Extraction des notes numériques
    - Nettoyage des commentaires
    - Identification des valeurs manquantes
    - Suppression des doublon
3. **Analyse de sentiment** :
    - Chargement du modèle distilcamembert-base-sentiment.
    - Chargement du *CamembertTokenizer*
    - Chargement du modèle ORTModelForSequenceClassification
    - Création du pipeline de classification
    - Application du modèle aux commentaires
4. **Visualisation :** Création de graphiques et d'analyses visuelles à l'aide des bibliothèques Pandas, NumPy, Matplotlib, seaborn et streamlit.

6- Rôle de chaque membre du groupe:

Jude KALENGAYI et Taous ZADDI: Scrapping des données.
Ouiza AIT RADI et Cheick Tidiane Sissoko : Prétraitement, nettoyage et classification.
Thelma LUM, Mai DAO et Cheick Tidiane Sissoko : Visualisation des données.
Jude KALENGAYI, Thelma LUM et Taous ZADDI : déploiement des résultats sur streamlit.

7- Outils :

- Pour réaliser le scrapping , nous avons utilisé **Python** ainsi que les bibliothèques suivantes :
    - **BeautifulSoup** : pour analyser les données HTML du site Décathlon.
    - **Selenium** : pour automatiser l'ouverture et l'interaction avec les pages du navigateur Chrome sur le site Décathlon et naviguer entre les pages.
- pour le prétraitement, nous avons utilisés python ainsi que les bibliothèques suivantes :

pandas : pour la manipulation et le nettoyage des données (extraction, gestion des valeurs manquantes, traitement de dates, suppression de doublons).

re : pour les opérations de nettoyage de texte avec les expressions régulières.

unidecode : pour la suppression des accents dans les textes.

datetime : pour la gestion et le formatage des dates.

- Pour l'analyse de sentiment nous avons utilisé Python ainsi que les bibliothèques suivantes :

  optimum avec ORTModelForSequenceClassification pour le chargement et l'optimisation du modèle avec ONNX Runtime.

- **transformers** avec CamembertTokenizer et pipeline pour le traitement de texte et l'interface avec le modèle de classification.

- Pour la visualisation de données , nous avons utilisé **Python** ainsi que les bibliothèques suivantes : pandas, plotly et datetime

  - Power BI : Création et publication des visualisations interactives.

  - Streamlit : Développement de l'application web avec intégration Power BI.

8 - Architecture cloud :

Pour déployer nos analyses et les graphes réalisés nous avons opté pour **Streamlit** qui est un framework Python open-source conçu pour les scientifiques des données et les ingénieurs en IA/apprentissage automatique qui permet de créer et de déployer facilement des applications interactives de données.

9 - Architecture Data :

Nous avons mis en place une architecture Data structurée en zones bronze, silver et gold. Cette approche nous permet de stocker les données brutes dans la zone bronze(Data Lake) , de les traiter et de les nettoyer dans la zone silver(Data Warehouse) , et enfin de conserver des données prêtes à l'analyse dans la zone gold(Data Mart) . Ainsi, nous pouvons suivre les différentes évolutions des données.

10 - Algorithme utilisé :

L'algorithme utilise le modèle distilcamembert-base-sentiment de Hugging Face pour analyser les commentaires en français. Le pipeline applique un tokenizer pour préparer les données, puis le modèle quantifié effectue la prédiction afin de classifier chaque commentaire en trois catégories : positif, négatif ou neutre, tout en attribuant un score correspondant pour évaluer l'intensité du sentiment.

11 - Avancement du projet :

Scraping des données :

Notre décision initiale s'est portée sur l'enseigne de vêtements SHEIN. Nous avons donc commencé le scraping sur ce site, mais nous avons remarqué que son architecture web était très complexe à exploiter.

Après réflexion, nous avons décidé de changer d'approche et de nous orienter vers l'enseigne Décathlon.

L'objectif initial de ce scraping était de sélectionner un article spécifique, en l'occurrence des chaussures de running, afin d'analyser les données associées. À partir de l'URL de la page produit, nous avons extrait tous les commentaires relatifs à cet article. Ces données ont ensuite été utilisées pour effectuer une analyse de sentiments sur le produit.

Prétraitement des données :

Dans cette partie, nous avons effectué plusieurs étapes de traitement sur les données afin de préparer et nettoyer les informations pour l'analyse.

Nous avons nettoyé la colonne *Auteur* en conservant uniquement la première partie de la chaîne avant le délimiteur | et en supprimant les espaces excédentaires. Nous avons converti la colonne *Date* en format datetime, géré les valeurs manquantes, au format année/ mois/jour. Nous avons extrait les valeurs numériques de la colonne note et les avons converties en entiers .

Nous avons appliqué une fonction pour nettoyer les commentaires en supprimant les accents, les caractères non alphabétiques, les espaces multiples et en convertissant le texte en minuscule. Nous avons vérifié les valeurs manquantes dans l'ensemble des données et supprimer les doublons pour garantir la qualité et la fiabilité des données.

Analyse de Sentiments et Classification:

Dans cette partie, nous avons réalisé une analyse des sentiments sur les commentaires. Cette analyse permet d'attribuer un score et de classifier les commentaires en trois catégories : **positif**, **négatif** ou **neutre**, en fonction des résultats prédits.

Pour cela, nous avons choisi d'utiliser **distilcamembert-base-sentiment** de **Hugging Face Hub**, une version adaptée à la classification de texte en français.

Afin d'utiliser ce modèle avec efficacité, nous avons chargé un tokenizer spécifique pour transformer les commentaires en une représentation compréhensible par le modèle.

Nous avons ensuite configuré un pipeline avec le modèle quantifié et le tokenizer. Ce pipeline permet de transformer automatiquement les commentaires en données exploitables par le modèle et d'effectuer l'analyse de texte.

Par la suite, nous avons appliqué l'analyse sur chaque commentaire. Deux nouvelles valeurs ont été générées pour chaque entrée :

1. La **note prédit**, calculée à partir de l'analyse du texte.
2. Le **score correspondant**, qui représente la probabilité de la prédiction.

Ensuite, nous avons classé les notes prédites selon le barème suivant :

- **1 étoile** : très négatif
- **2 étoiles** : négatif
- **3 étoiles** : neutre
- **4 étoiles** : positif
- **5 étoiles** : très positif

Visualisation des données avec Power BI

- **Importation des données :** Le fichier `prediction.csv` a été chargé dans Power BI avec des traitements dans Power Querry.
- **Création des visuels :** Nous avons créé une courbe temporelle, un histogramme pour la distribution des sentiments, et un graphique comparant les notes réelles et prédites et un words cloud.

Développement et déploiement du dashbord sur Streamlit

- **Interface utilisateur :** Une interface simple avec boutons, champs de saisie, et graphiques interactifs a été conçue.

12 - Schéma d’architecture :