Web Scraping for Assessing Customer Satisfaction on an E-Commerce Site

Decathlon Products Customer Review

Presented by :
Elaa Marco
Lina Smiri
Siwar Jerbi
Chaima Chammaa

IT-360 Information Assurance and Security

Tunis Business School

May 2025

Introduction

This project focuses on developing a Product Review Analyzer through web scraping techniques. In today's data-driven world, web scraping has become an essential tool for businesses to:

- Gather customer feedback
- Monitor competition
- Make data-informed decisions

Our project aims to extract and analyze product reviews from e-commerce sites to assess customer satisfaction levels

Main Concepts

- **Definition of Web Scraping :**

is the process of automatically extracting data from websites. This technique is commonly used to gather large volumes of data from the internet for analysis, monitoring, or integration into other systems

- **Functional Flow :**

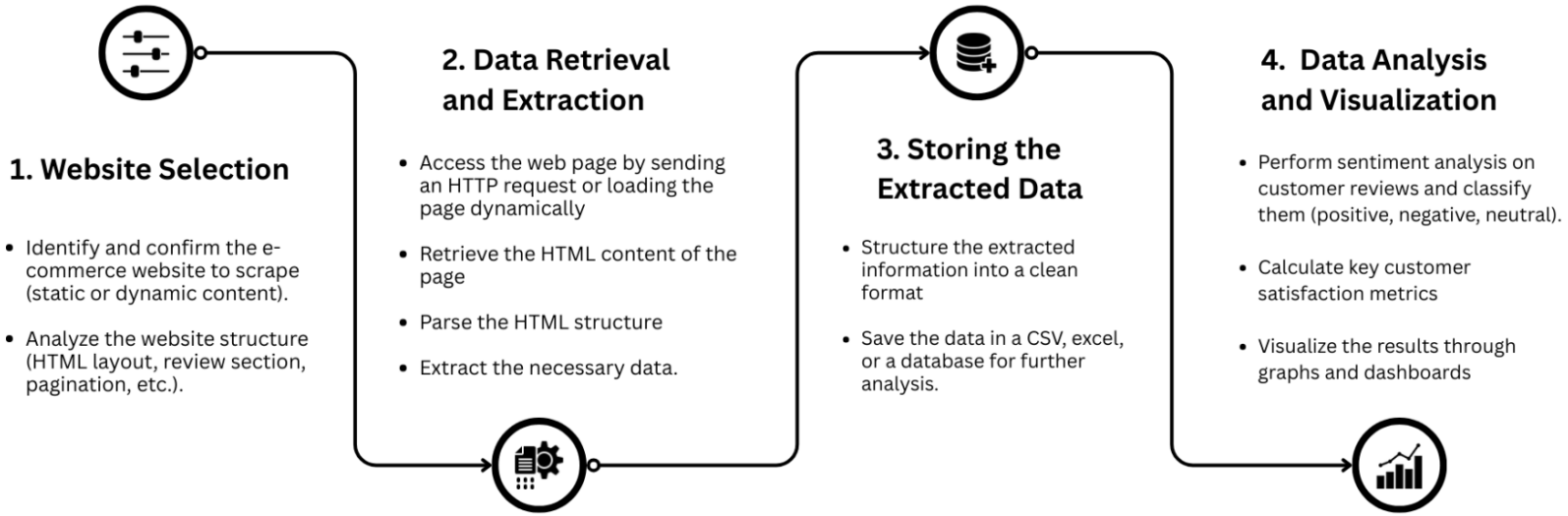

Figure 1: Functional Flow Diagram

Tools and Libraries

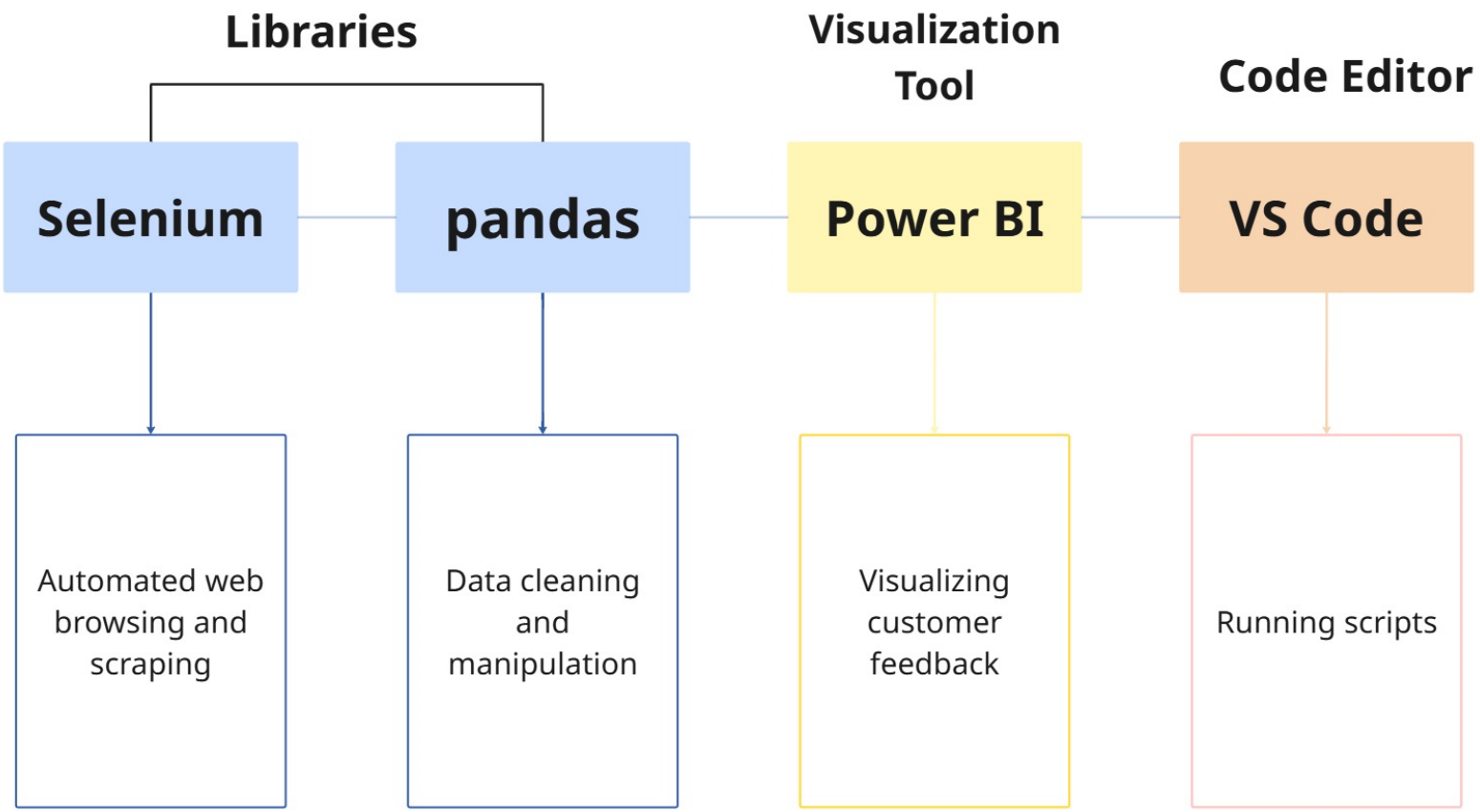

Figure 2: Tools and Libraries Used

Technical Framework

Our Framework consists of **3 core steps** from the automation of the process of collecting customer feedback to analyzing the satisfaction levels through graphical representation :

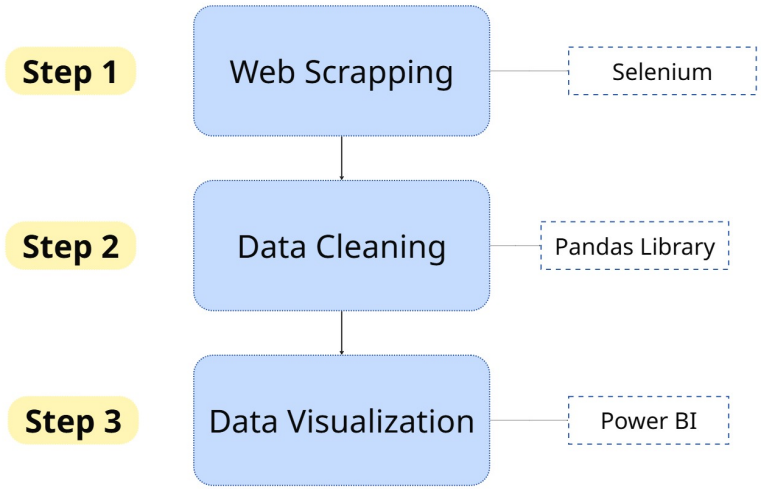

Figure 3: The Main Steps

Web Scrapping

Purpose of the Script:

This script automates the scraping of customer reviews from Decathlon Tunisia product pages using Selenium. It extracts review data, prevents duplicates, and saves the output to a CSV file for each product.

Component	Description
Scraping Tool	Selenium WebDriver with Chrome, using automated driver installation
Target Website	Product pages on Decathlon Tunisia
Data Extracted	Rating, Date, Title, Content, Reviewer, Verified status, Brand response
Output	One CSV file per product with extracted review data
Error Handling	Uses try-except blocks to prevent crashes from missing elements or timeouts
Max Reviews Limit	Up to 3000 reviews per product; pagination capped at 13 pages
Duplicate Check	Avoids duplicates using a hash key of title, content, and date
Language and Headers	French-language interface via browser settings and user-agent
File Naming	CSV named using a cleaned product name and current date

Table 1: Overview of the Decathlon Review Scraper Script

Data Cleaning

```python
import pandas as pd
import os
import glob

# Set your folder path
folder_path =r"C:\Users\Dell\OneDrive - Ministere de l'Enseignement Superieur et de la Recherche Scientifique\Bureau\CSR files sec project"  # Change this to your folder path
output_folder = os.path.join(folder_path, "cleaned")
os.makedirs(output_folder, exist_ok=True)

# Loop through all CSV files in the folder
csv_files = glob.glob(os.path.join(folder_path, "*.csv"))

for file_path in csv_files:
    try:
        # Load data
        df = pd.read_csv(file_path)

        # Drop rows with missing essential data
        df = df.dropna(subset=['Content', 'Date'])

        # Convert Date to datetime format
        df['Date'] = pd.to_datetime(df['Date'], format="%d/%m/%Y", errors='coerce')
        df = df.dropna(subset=['Date'])  # Drop rows with invalid dates

        # Fill missing values
        df['Title'] = df['Title'].fillna("No Title")
        df['Brand Response'] = df['Brand Response'].fillna("No response")

        # Create Year-Month column
        df['YearMonth'] = df['Date'].dt.to_period('M')

        # Extract Country
        df['Country'] = df['Reviewer'].str.extract(r',\s*(\w+)$')

        # Reset index
        df = df.reset_index(drop=True)

        # Save cleaned file
        filename = os.path.basename(file_path)
        cleaned_path = os.path.join(output_folder, f"cleaned_{filename}")
        df.to_csv(cleaned_path, index=False)

        print(f"✅ Cleaned: {filename}")

    except Exception as e:
        print(f"❌ Error with {file_path}: {e}")
```

Figure 4: Cleaning Script

Step	Description
Load CSV	Read raw CSV file using `pandas.read_csv()`
Drop Missing	Remove rows where `Content` or `Date` is missing
Convert Dates	Convert `Date` column to datetime format with error coercion
Filter Invalid Dates	Drop rows where date conversion failed (NaT values)
Fill NA Values	Replace missing `Title` with "No Title" and `Brand Response` with "No response"
Add Year-Month	Create a `YearMonth` column from the cleaned date
Extract Country	Extract country name from the `Reviewer` field using regex
Reset Index	Reset row indexing after all modifications
Save Clean File	Export cleaned data as a new CSV in a subfolder `/cleaned`

Table 2: Summary of Cleaning Steps

Data Visualization

After collecting and cleaning customer reviews from Decathlon product pages, we used data visualization to uncover key satisfaction trends. Through dynamic dashboards in Power BI, we highlighted common themes, average ratings, and frequency of verified reviews—making it easier to interpret large volumes of unstructured feedback.

All technical details and scripts for the different steps are available in the GitHub Repository

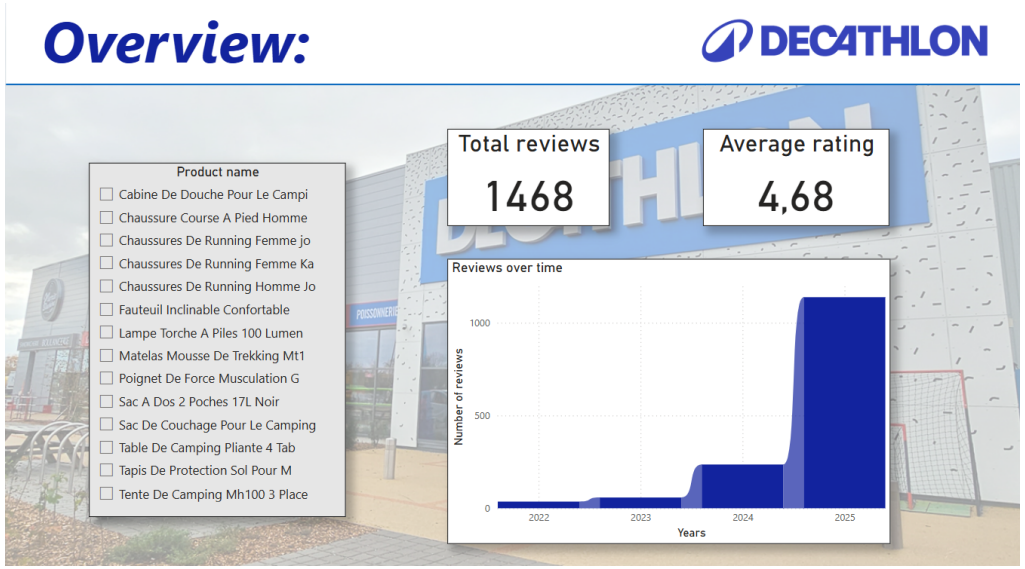

Figure 5: Overview Page

⇒ This overview provides a high level summary of the reviews. The total number of reviews is 1468, with an average rating of 4.68. The "Reviews over time" chart shows a significant surge in reviews in 2025, indicating a recent increase in customer feedback.

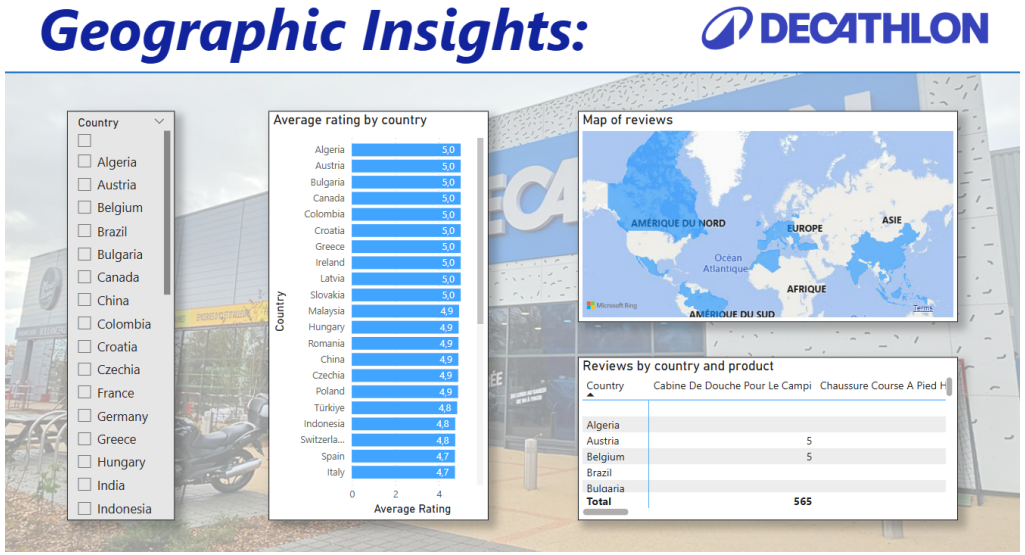

Figure 6: Geographic Insights

⇒ This Figure highlights the geographic distribution of reviews and average ratings. It seems many countries have a perfect average rating of 5.0, while others like Italy and Spain have slightly lower averages. The map visually confirms the global reach of the reviews, and the table on the bottom right shows the number of reviews per country for specific products, with a total of 565 reviews listed.

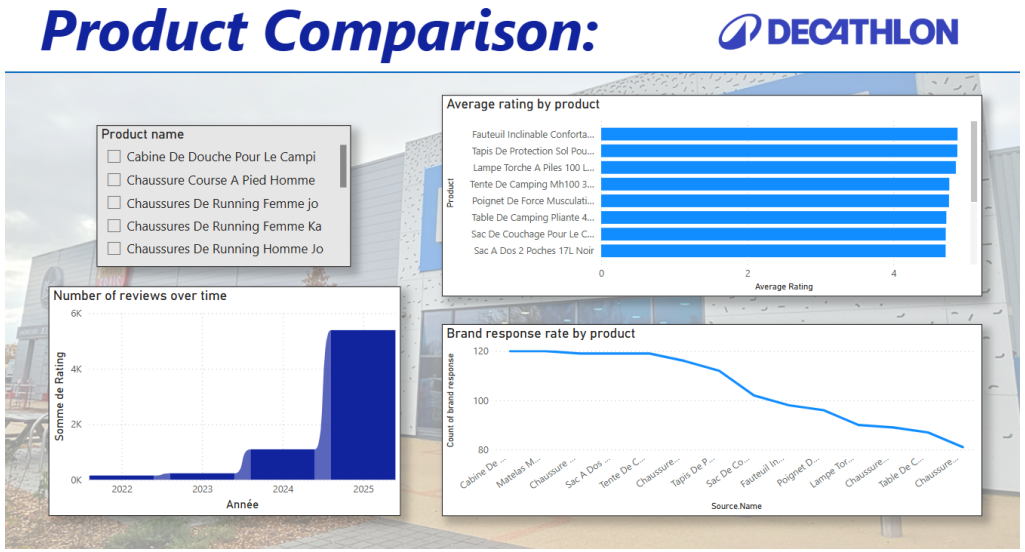

Figure 7: Product Comparison

⇒ This Figure allows for a comparison of products based on average rating and the number of reviews over time. The "Brand response rate by product" chart indicates varying levels of engagement across different products, with some products receiving significantly more brand responses than others. This view helps identify top-rated products and those with the most feedback.

Review Content Analysis: DECATHLON

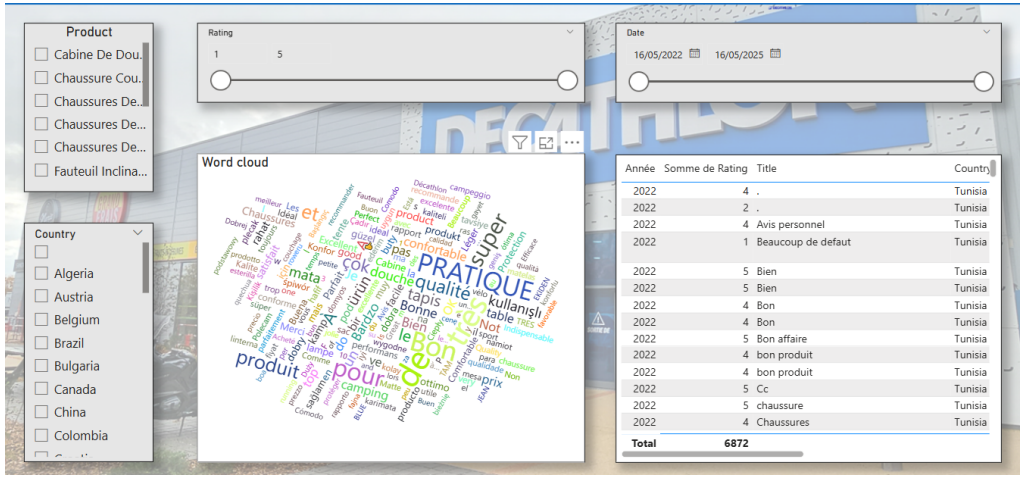

Figure 8: Review Content Analysis

⇒ This Figure shows the content of the reviews, showcasing a word cloud with prominent positive terms like "pratique," "bon," and "super," suggesting generally favorable feedback. Filtering options for product, rating, and date are available. The table on the right provides specific review details, including the year, rating, title, and country, with a total of 6872 reviews in this filtered view.

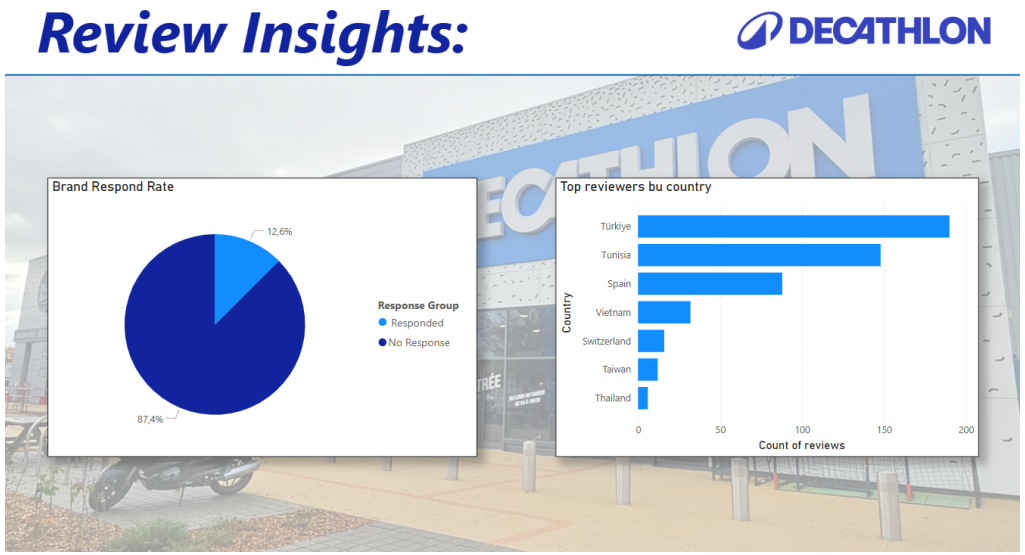

Figure 9: Review Insights

⇒ This Figure focuses on brand responsiveness and the origin of top reviewers. The brand response rate is quite low, at only 12.6%. Turkey appears to have the highest number of reviewers, followed by Tunisia and Spain. This suggests that customer engagement through responses could be improved, and marketing efforts might consider the countries with the most active reviewers.

This project demonstrates the effectiveness of web scraping and data visualization in capturing and understanding customer sentiment. By automating data collection from Decathlon and interpreting it visually, we gained actionable insights into customer satisfaction trends, helping businesses refine their customer engagement strategies.

Thank You For Your Attention!