Institution:
Tunis Business School

Subject:
Information Assurance and Security

Project

Market Assessment of Electric Vehicles
Using Web Scraping

Authors

Salma Azouzi, Eya Maalej, Sarra Jebali,
Meriem Houissa, Kmar Abessi

Professor: Mrs. Manel Abdelkader
Academic Year: 2024–2025

Major: Business Analytics **Minor:** IT

Summary

This project transforms scattered online noise into actionable insights to empower prospective electric vehicle (EV) buyers. By automating data collection from official websites, online marketplaces, customer review platforms, and social media, it replaces manual research with a smart web scraping system. Using data analysis and data mining algorithms, it decodes public sentiment, highlights common complaints, and benchmarks key EV features such as battery life, charging time, and performance. An integrated alert system also notifies users of new model releases and exclusive deals, giving buyers a confident, informed edge in choosing the right EV. The result is an intelligent, data-driven platform that presents real-time pricing trends, feature comparisons, and brand reputations through a user-friendly dashboard.

Contents

1 Introduction

1.1 Topic

This project provides a practical, data-driven assessment of electric vehicle (EV) options for potential buyers. It leverages automated data collection and sentiment analysis to guide decision-making.

1.2 Why Electric Vehicles?

Electric vehicles have gained popularity as battery technology and design have improved. The EV market is now diverse and competitive, making buying decisions more complex. Consumers must weigh performance, pricing, and brand reputation.

1.3 Project Goal

To automate the EV purchasing process by:

- Comparing features and prices across brands
- Analyzing customer sentiment and frequent complaints
- Tracking deals and new model releases

1.4 Scope and Deliverables

- **Data Extraction:** From social media, review platforms, and official EV websites.
- **Sentiment Analysis:** To capture public opinion and identify trends/issues.
- **Alert System:** Notifies users of new models or discounts.
- **Visual Dashboard:** Presents trends, comparisons, and insights in an interactive format.

2 Theoretical Background

2.1 Web Scraping

- **Definition:** Automated extraction of unstructured HTML data and conversion into structured formats (CSV, JSON).
- **Components:**
  - Crawler: Navigates URLs
  - Parser: Extracts/cleans HTML
  - Scheduler & Pipeline: Organizes and stores data
- **Mathematical Background:**
  - Web scraping can be modeled as stochastic sampling from a large, dynamic dataset (the web).

  - It represents web pages as a set P, where the scraper selects a subset $S \subset P$.
  - After making an HTTP request to a selected URL, a response is received in the form of an HTML document. The HTML document can be represented as graph $G(V, E)$ with:
    * V: HTML tags (e.g., ¡div¿, ¡a¿)
    * E: Parent-child edges
  - The scraping bot employs traversal algorithms such as Depth-First Search (DFS) or Breadth-First Search (BFS) to extract structured data from raw HTML content. It can also handle pagination, dynamic loading, and CAPTCHAs using headless browsers (e.g., Puppeteer or Selenium).

2.2 Natural Language Processing (NLP)

- NLP is a subfield of artificial intelligence that focuses on the interaction between computers and human language. It enables machines to understand, interpret, and generate human language.
- Use in this project: Analyzing customer reviews, tweets, and forum posts to determine brand sentiment or perceived value.
- Techniques:
  - Rule-based (e.g., VADER)
  - Machine Learning (Naive Bayes, SVM)
  - Deep Learning (e.g., BERT)
  - Mathematical Basis:

$$\text{Polarity Score} = \frac{\sum word\ polarities}{word\ count}$$

2.3 Sentiment Analysis

- Sentiment analysis is the computational task of automatically determining the emotional tone behind a series of words.
- Use in this project: Understand public perception of brands.
- Techniques:
  - Text Vectorization:
    * TF-IDF(Term Frequency-Inverse Document Frequency) is a technique used to transform text into numerical vectors.
    * The TF-IDF value for a term t in a document d is given by: $TF\text{-}IDF(t, d) = TF(t, d) \times \log(N/df(t))$
  - Classification Model:

* Logistic Regression: is commonly used for binary sentiment analysis (positive vs. negative)
  $y = 1/(1 + e^{-(w^T x+b)})$ where:
  x: vectorized text
  w: weight vector
  b: bias term
  y: predicted probability of the positive sentiment.
* Softmax Regression: can be used for multi-class sentiment classification (positive, negative, neutral)

$$P(y = j \mid x) = \frac{e^{\theta_j^T x}}{\sum_{k=1}^{K} e^{\theta_k^T x}}$$

  Where K is the number of sentiment classes.

- The bot will capture textual data and pipe it into a sentiment classifier using tools such as VADER, TextBlob, or fined-tuned BERT models.

2.4 Feature Benchmarking

Feature benchmarking is the process of comparing technical specifications and product features across different competitors to evaluate performance and value.

- Use in this project: Comparing battery, range, acceleration, charging time
- Steps:
  - Preprocessing:
    * Min-Max: $x' = (x - min(x))/(max(x) - min(x))$
    * Z-score: $x' = (x - \mu)/\sigma$
  - Scoring: Score $= \sum w_j x_j$
  - Similarity:
    * Euclidean Distance:

$$d_{\text{Euclidean}}(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

    * Manhattan Distance:

$$d_{\text{Manhattan}}(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

  - Clustering: K-means which groups EVs with a similar features patterns.
  - Dimensionality Reduction: PCA (Principal Component Analysis), which reduces multiple features into a few summary axes for visualization and analysis.
  - MCDM: TOPSIS which ranks EVs by their closeness to an ideal solution.

2.5 Python Libraries

- Requests: Simplifies HTTP requests to retrieve web pages.
- BeautifulSoup:Parses HTML and XML documents to navigate the document tree and extract desired data.
- Selenium: A browser automation tool that helps scrape content from dynamic websites.

3 Functional Flow

3.1 Platforms and Data Points

- Official Sites: Specs (Features, Battery, Acceleration, Charging Time) and pricing
- Marketplaces: Trends, offers (e.g., CarGurus, EV.com, Autotrader), and availability.
- Social Media: Hashtags, mentions, buzz metrics (views, likes), user sentiment (X, Reddit, etc.)
- Review Sites: Customer reviews, ratings, complaints (e.g., Trustpilot, Formus)
- Deal Trackers: Promotions, discounts, model releases (e.g., dealership pages)

3.2 Data Collection

- Modular scrapers tailored to static/dynamic sites.
- Selenium used for JS-rendered content (e.g., YouTube)
- Employ delays, proxies, and user agents to avoid bans.

3.3 Data Preprocessing

- Deduplication
- Unit and currency normalization
- Structuring by type: price, sentiment, features

3.4 Analysis Workflow

- Pricing Trends
  - Average price per model, region
  - Outlier detection
- Benchmarking
  - Normalize specs, compute similarity
  - clustering of comparable vehicles

- Sentiment Analysis
  - Classify text sentiment (positive, neutral, negative)
  - Track changes over time or by model
- Buzz Tracking
  - Count mentions/hashtags
  - Analyze engagement (views, comments)

3.5 Visualization and Reporting

- Charts, timelines, heatmaps
- Alerts: discounts, drops
- Sample Reports: "Best Value EV Under $40K$", "Top 3 in Range and Battery Life", "Tesla vs. BYD"

4 Existing Tools & Solutions

Tool	Features	Advantages	Limitations
BeautifulSoup	HTML parsing	Easy, fast for small tasks	Poor on dynamic content
Scrapy	Scalable framework	Modular, fast	Complex setup
Selenium	Browser automation	Dynamic content support	Slow, heavy
Octoparse	Visual tool	Cloud-based, easy	Less flexible
ParseHub	AJAX support	UI-friendly	Not customizable

5 Advantages

- Scalability: Fast, large-scale data extraction.
- Customizability: Tailored scraping and alerting.
- Depth of Insight: Captures granular user sentiment and product comparisons.
- Low-cost: Open-source tools; ideal for startups or students.

6 Challenges & Limitations

- Maintenance: Scrapers break when websites change.
- Legal risks: Risks is scraping violates terms or robot.txt
- Data quality: May include noise, bias, sarcasm, or duplicates.
- Performance: Large-scale scraping/analysis can be resource-heavy.
- Platform barriers: Restricted APIs or scraping protections (e.g., Instagram, LinkedIn).