Dynamic E-commerce Data Extraction

Team Details
1. B.LAXMI PRIYA (20EG105306)
2. NAGHMA MULLA (20EG105332)
3. T.DINESH REDDY (20EG105350)
4. P.SRIYA REDDY (20EG105722)

Project Supervisor
G.KIRAN KUMAR
ASSISTANT PROFESSOR

Department of Computer Science and
Engineering

Introduction

In the fast-paced world of online shopping, the demand for accurate and timely product information is paramount. The abundance of e-commerce websites poses a unique challenge—extracting precise data efficiently. To address this challenge, our project, titled "Dynamic E-commerce Data Extraction," endeavors to develop a sophisticated system that goes beyond conventional methods. This initiative aims to revolutionize the online shopping experience by seamlessly providing reliable product details, adapting to ever-changing website structures, and prioritizing user experience.

Problem Statement

Acquiring accurate product details from online shopping sites is complex. Our project aims to create an efficient system using advanced technology for text comprehension and image recognition. This system will adapt to website changes, improving the online shopping experience by swiftly providing reliable information. Creating an efficient e-commerce data extraction system involves diverse components. Web scraping and API integration gather data, while text comprehension and image recognition process information. Adaptation strategies and reinforcement learning ensure system flexibility.

Proposed Method

The proposed method for addressing the challenges in dynamic e-commerce data extraction involves a multi-faceted approach:

Website Snapshots: Capture visual representations of e-commerce websites to analyze layout structures and visual elements**.**

Object Detection (YOLO): Swiftly identify and locate objects, with a focus on product-related elements, using the YOLO object detection algorithm.

Text Extraction (Tesseract OCR): Utilize Tesseract OCR to extract textual information from identified objects, converting text within images to machine-readable data.

Data Processing: Process and organize extracted data to ensure cleanliness, filter out noise, and structure information for subsequent analysis.

Bert Transformation Model: Leverage the tBert transformation model to establish semantic relationships between textual information and products, enhancing contextual understanding.

Proposed Method

Transfer Learning: Apply transfer learning in deep learning models to adapt to changes in website structures, ensuring sustained efficiency in data extraction without manual intervention.

Feature Engineering: Enhance the dataset by crafting new features that contribute to the overall understanding of the data, providing more relevant and valuable information.

Data Analysis: Analyze the processed data for patterns, insights, and valuable information, extracting actionable intelligence for further use.

Report Generation: Generate a comprehensive report based on the analyzed data, summarizing key insights, trends, and findings for informed decision-making.

Specifications for an Improved User Experience: Give users the ability to define particular data points that they are interested in and then tailor the extraction process to address their specific needs.
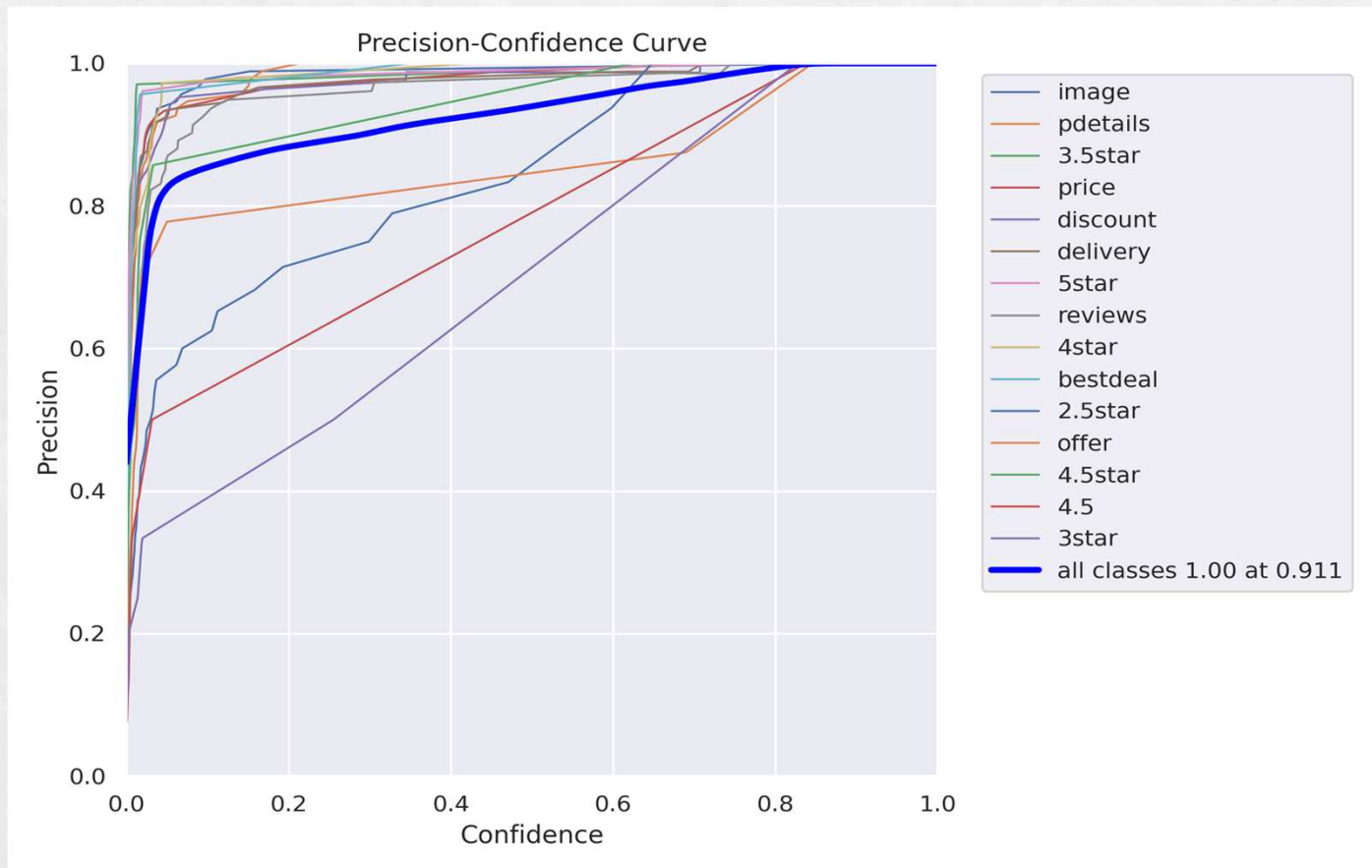
Experiment Environment

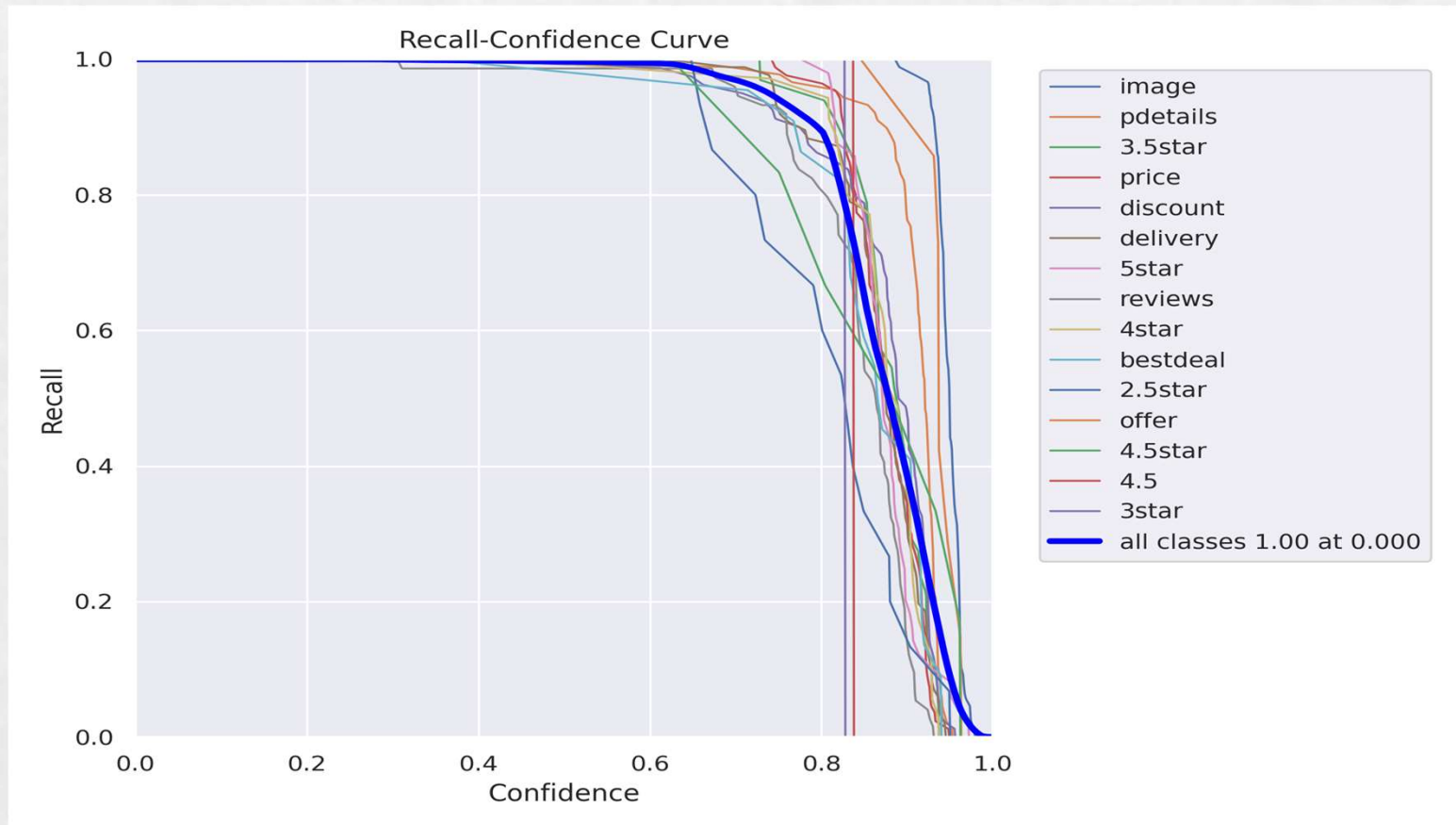**Tool, language, data set … etc**

Operating system: Ubuntu
Frameworks & libraries: TensorFlow, Flask, Pytorch, NumPy, pandas and NLTK
Additional tools: jupyter notebook, google colab and VS code

Department of Computer Science and
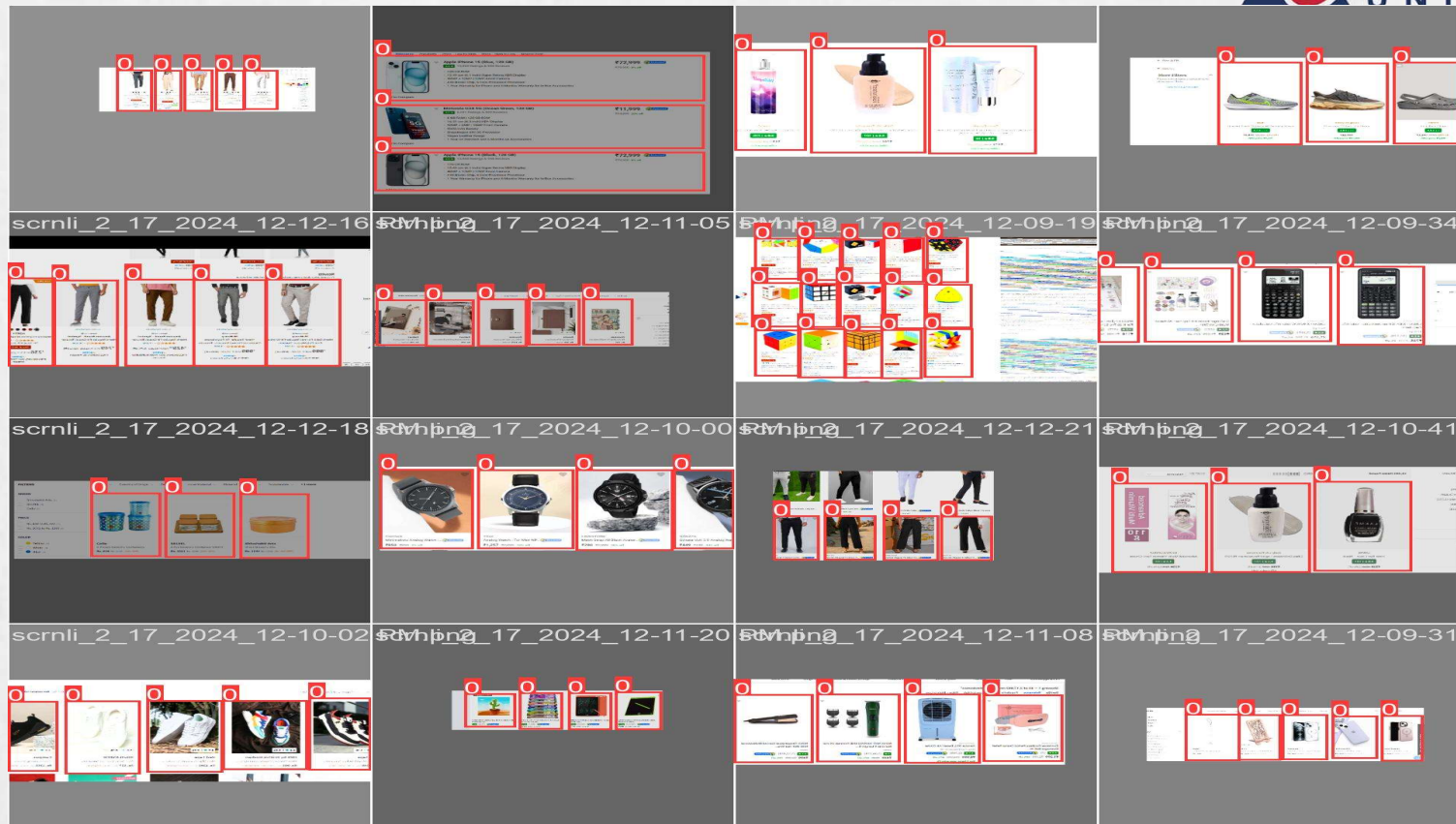Engineering

Experiment Screen shorts

Department of Computer Science and Engineering

Experiment Screen shorts

Department of Computer Science and Engineering

Experiment Screen shorts

Experiment Results
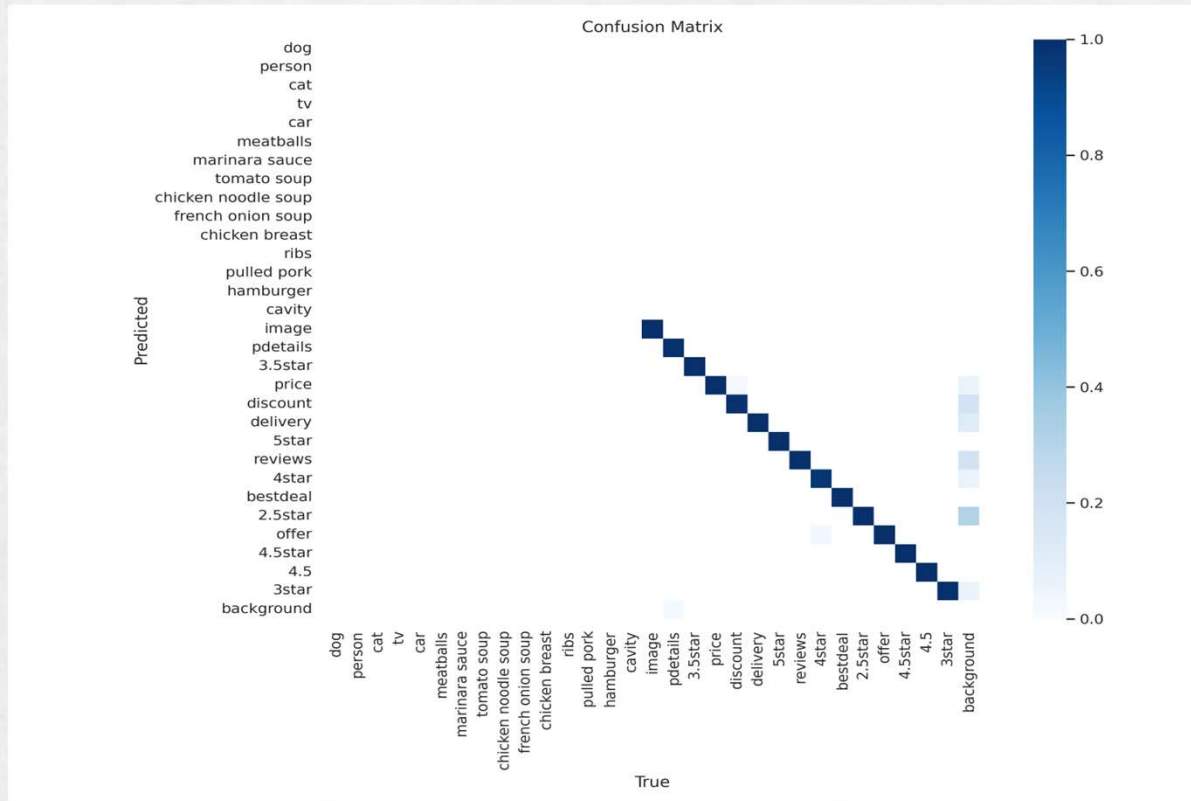
Department of Computer Science and Engineering
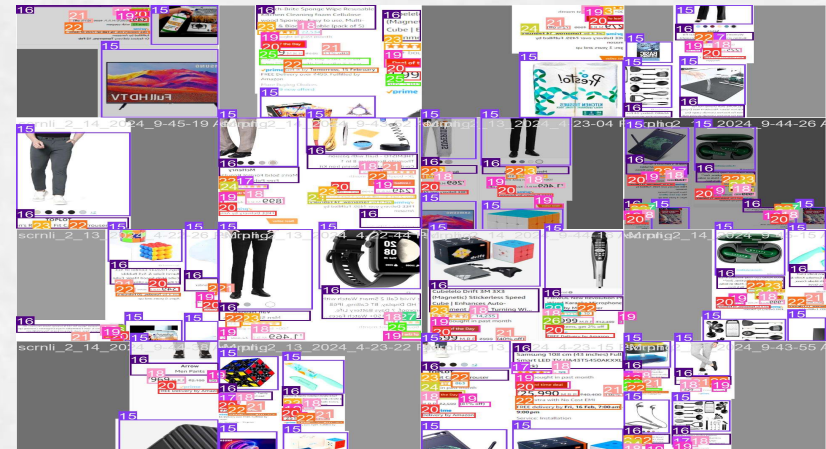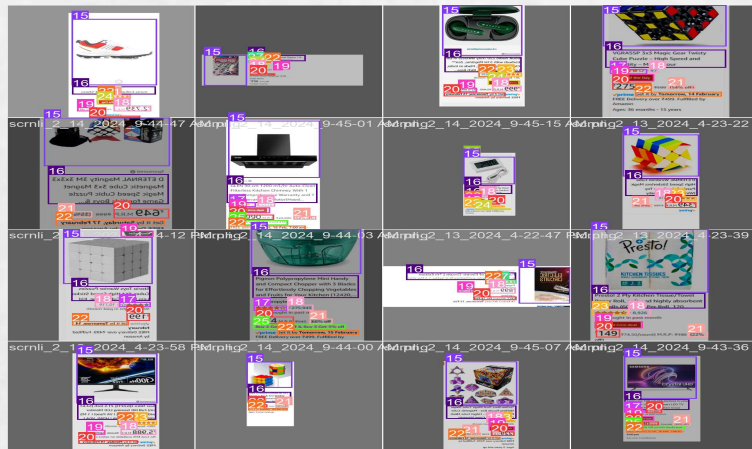
Experiment Results

Experiment Results

Department of Computer Science and Engineering
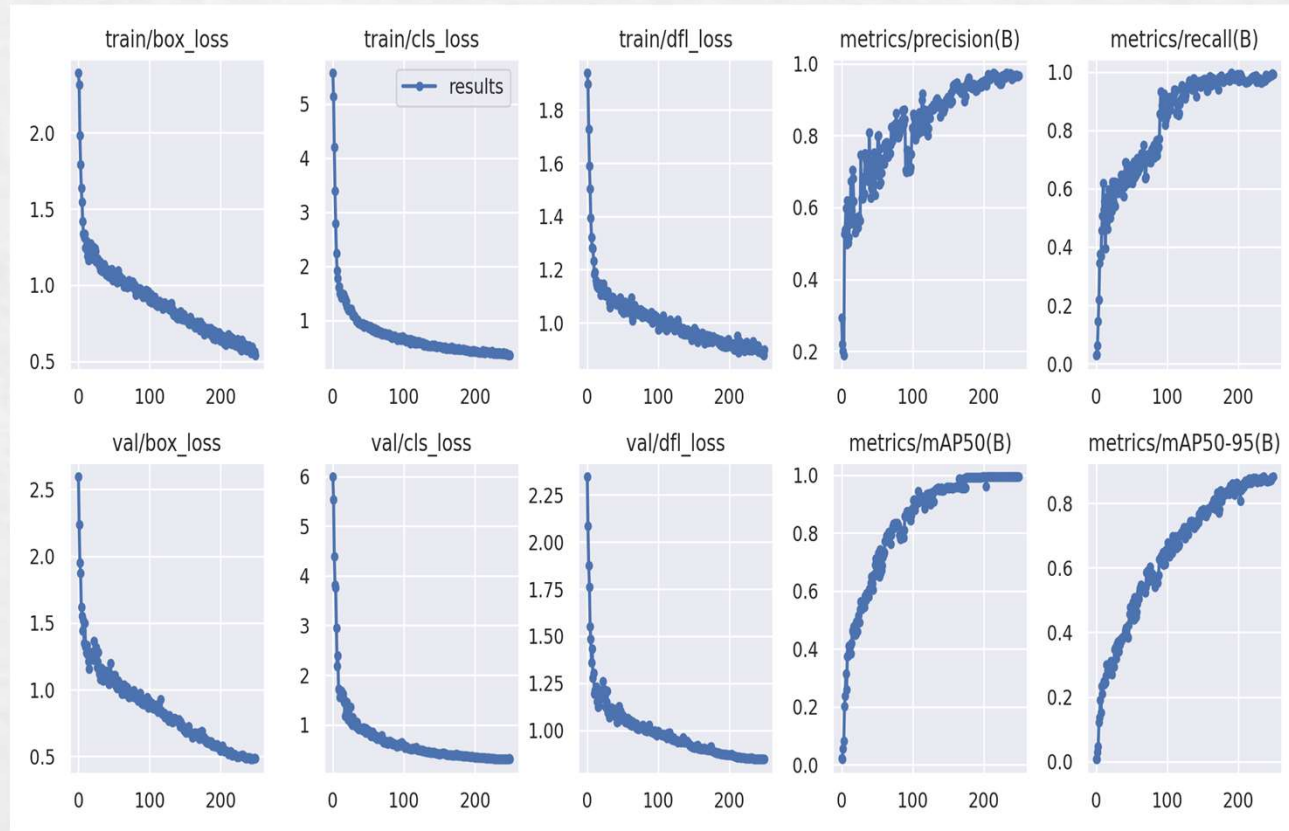
Finding

This project successfully built a system to automatically extract structured product details, pricing, and reviews from e-commerce websites, enhancing trust and satisfaction among users. It improves inventory management, order fulfillment, and targeted marketing campaigns. While the system struggles with very similar items, expanding the data used to train it and incorporating computer vision techniques can enhance accuracy, making it a powerful tool for extracting e-commerce data. Overall, this project focuses on enhancing dynamic data extraction in e-commerce to improve precision, drive engagement and conversions, adapt to market fluctuations, and differentiate platforms for sustainable growth and customer loyalty.

Justification

Dynamic data extraction in e-commerce is indispensable for maintaining relevance and accuracy in product listings, pricing, and availability. By leveraging advanced algorithms and technologies, such as web scraping and API integrations, businesses can ensure that their platforms deliver up-to-date information to customers. This precision not only enhances trust and satisfaction among users but also reduces the likelihood of errors or discrepancies that could negatively impact the shopping experience. Furthermore, by continuously monitoring and updating data in real-time, e-commerce platforms can adapt swiftly to market fluctuations, ensuring that customers always have access to the most relevant and reliable information.

Moreover, enhancing dynamic data extraction capabilities in e-commerce contributes significantly to improving user experience. With precise and timely information, businesses can personalize product recommendations, promotions, and notifications based on individual preferences and behavior. This level of customization fosters a sense of connection and relevance for users, ultimately leading to increased engagement and conversion rates. By prioritizing accuracy and user-centricity in data extraction processes, e-commerce platforms can differentiate themselves in a crowded market and drive sustainable growth and customer loyalty.