CoreID JOURNAL

Vol. 3 No. 1, March 2025 (37-46)
ISSN: 2987-6990
DOI: https://doi.org/10.60005/coreid.v3i1.90

Analysis Of E-Commerce Product With Web Scraping Technique

Siti Jahro Maulidiyah[1], Asep Indra Syahyadi[2]
[1]Department of Informatics Engineering, Faculty of Science and Technology, UIN Sunan Gunung Djati, Bandung, Indonesia
[2]Department of Informatics Engineering, Faculty of Science and Technology, UIN Alauddin, Makassar, Indonesia

Article Info

Article history:

Received January 15, 2025
Revised February 24, 2025
Accepted March 19, 2025

Keywords:

Web Scraping
E-commerce
Data Extraction
API Integration
Market Analysis
Data Visualization
West Java Statistic

ABSTRACT

This research aims to implement a web scraping system to automatically extract product data from the e-commerce platform Bukalapak, with the goal of supporting statistical analysis at the Central Bureau Statistics (BPS) of West Java Province. The system utilizes a combination of API access and automation tools such as python, executed in the Google Colab cloud environment. Through this method, 74,796 product records were successfully collected, encompassing information such as product names, prices, categories, customer reviews, stock levels, and seller locations. The data was then processed and visualized using bar charts and histograms to analyze market trends, price distribution, and consumer behavior across regions in West Java. The results show that most products fall within affordable ranges, with certain categories like electronics and personal care dominating in volume. The scraping approach proved to be an efficient and scalable solution for acquiring real-time market data, supporting BPS in evidence-based decision-making and policy formulation.

Corresponding Author:

Siti Jahro Maulidiyah
Department of Informatics Engineering, Faculty of Science and Technology, UIN Sunan Gunung Djati Bandung, Indonesia
Email: szmaulidiyah15@gmail.com

1. INTRODUCTION

Today, the internet has become an important necessity. With the internet, the development of technology and the need for information has increased[1]. Market place is a website that can connect directly between buyers and sellers through internet technology. With the website provided by the marketplace, traders and buyers do not have to meet directly in buying and selling activities. The field of e-commerce itself is quite broad, usually including distribution services, sales, purchases, services, maintenance, and product services that all take place in electronic systems such as the internet or other computer networks[2]. The marketing process in the marketplace continues to develop along with technological developments into a digital marketing process such as utilizing e-commerce. Economic activity aims to improve the economic performance of the people and the quality of life as a whole[3]. Websites are now the primary means for Companies, organizations and individuals to communicate with their online audience in this digital age that is constantly evolving. Program testing is an unnoticed but very important process, hidden, behind elegant designs and complex functionality. Very important in the software development cycle is website program testing. This is a step to ensure that the website can offer a positive user experience in addition to good functionality. Program testing benefits businesses by reducing the likelihood of failure, increasing client confidence, and ensuring compliance with and industry standards. The central Bureau of Statistics (BPS) of West Java Province needs data on e-commerce products for various needs, because of the importance of e-commerce in understanding the dynamics

of the digital economy which is very trendy at this time, for example, such as monitoring the development of the digital economy, monitoring prices and inflation of Digital Products, the prices of goods on e-commerce platforms often change so it is difficult to monitor price stability directly, as well as mapping E-commerce-based UMKM. The provincial government wants to know how UMKM utilizes e-commerce platforms to increase income, but it is difficult to monitor the development of UMKM without detailed data. E-commerce is an online shopping activity using the internet network and its transactions through digital money transfers brings great business opportunities and revenue growth. It provides a choice of shopping methods for people by no longer using traditional methods such as coming directly to the store as the main way[1]. Web scraping is the process of taking a semi-structured document from a web page, to extract only certain data from that page [4]. The process of web scraping from the internet is basically divided into two sequential steps, namely finding the web whose data will be extracted and then extracting the data/information needed from the web[4]. API provided by a particular website to load data on the website? Twitter and Amazon are examples of websites that provide API services where you can access the API to get the data you need. However, using API has several drawbacks such as. Web scraping, also known as web data extraction, web harvesting, or screen scraping, aims to extract information from websites into understandable data such as spreadsheets database, or Comma-Separated Values (CSV) files [5][6]. There are various web scraping methods that can be used, ranging from manual methods such as copy-pasting to automated methods. The copy-pasting method is a simple and easy-to-use method, done by opening a browser and manually copying data to paste it into other media[7]. Processing needs are adjusted to the needs of the data analysis itself. Before processing, the dataset or collection of data to be processed must be available first. There are two types of datasets, namely: (1) private datasets, which are datasets taken from the organization/company/agency that owns the data, for example, hospital data, banks, schools, companies, and so on; (2) public datasets, which are datasets taken from public repositories. In addition to this, there are also data whose retrieval must be done by crawling, for example Twitter data. Digital transformation has created a new platform for doing business and buying and selling, namely e-commerce. E-commerce or electronic commerce is a transaction or buying and selling activity using electronic media facilities, in this case the internet. E-commerce is also a medium to market and promote products[8].

The purpose of this research is to implement a web scraping system that is able to extract product data automatically, collect, retrieve, and store product data from e-commerce pages, including information such as product name, price, description, stock, customer reviews, product location and category, and other relevant information. The data collected is expected to support statistical analysis related to market trends, consumption patterns, price distribution, and availability of goods by region in West Java Province.

2. METHOD

Websites are now the primary means for Companies, organizations and individuals to communicate with their online audience in this digital age that is constantly evolving. Program testing is an unnoticed but very important process, hidden, behind elegant designs and complex functionality. Program testing benefits businesses by reducing the likelihood of failure, increasing client confidence, and ensuring compliance with relevant laws and industry standards.

a) Data Collection Stages

The initial stage begins with creation of a list of keywords and location codes stored in an Excel file. This file becomes a reference in the Scraping process so that the system can retrieve data based on a combination of product names and regions. The authentication process to Bukalapak API is done using access token inserted in the header of each HTTP request. This token ensures that only authorized users can access data from the API. The main function in the scraping_items, which automatically sends product data request based on specified keywords and locations. Parameters sent include the number of pages, the number of results per page, and other search information. Each response from the API is processed into structured data such as product name, price, category, location, rating, number of reviews, and product URL.

b) Error Handling and Request Limits

To avoid rate limits and IP blocking, the python script includes a timeout using time.sleep() on every API request. The system is also designed to handle possible errors or empty responses from the API server, so that the scraping process continues to run stably.

c) Data Storage and Visualization

The scraped data Is stored in CSV and Excel formats to facilitate the analysis process. Furthermore, the data is visualized using bar charts and histograms, including :

1. Distribution of initial price and current price
2. Number of products per category
3. Distribution of products by region (location code)

4. Average rating and number of reviews per product.

d) Python
Python is an object based programming language that can be interacted with interactively. This language has high-level data structure. Python is an interpretive programming language that has many functions, and is designed with a focus on clarity and ease of understanding of the code. Python is considered a language that combines the power and clarity of code syntax. Python programming is specifically designed to make it easier for programmers to create programs with time efficiency, ease of development, and compatibility with the system. Python can be used to create standalone applications or script programming [9].

e) Application Programming Interface (API)
allows developers to integrate two parts of one application with different applications simultaneously, which aims to speed up the development process by providing separate item functions so that developers do not need to create similar features [10].

f) Framework
Framework is a software that makes it easy for programmers to create a web application that has various functions such as plugins, and concepts to form a certain system so that it is neatly arranged and structured. Using a framework does not mean that you will be free from coding. As a framework user, you are required to use the functions and variables in a framework that is being used [11].

g) Website
Website can be defined as a collection of pages containing digital data information in the form of text, images, animations, sounds, and videos or a combination of all of them provided through an internet connection so that it can be accessed and viewed by everyone around the world A website contains all the web pages contained in a domain that contain information. A website is usually built on many interconnected web pages [12].

h) Browser
A Browser is software used to access and display web content. Browsers allow users to navigate, read, and interact with web pages [13]. Browsers work by interpreting and executing HTML, CSS, and JavaScript code to visually display web content to users [12].

This process helps in understanding market trends and consumption patterns of people in West Java Province, and can be used by the Central Bureau of Statistics (BPS) to support data-based decision making.

3. RESULTS AND DISCUSSION

Web Scraping is a program or automated code that can visit and search for information according to keywords on a website. Web Scraping travels on the web and serves to collect all information about a web page and index it into a database[14]. Information about the web page is obtained from the words which are usually used as keywords to find web pages. This process of collecting information about web pages from websites are called Web Scraping[15]. A program language is a standardized set of instructions for instructing a computer to perform certain functions. It is a set of syntax and semantic rules used to define computer programs. It allows a programmer to specify exactly which data computer will process, how this data will be stored/transmitted and exactly what type of steps will be taken in various situations [16]. Framework is a software that makes it easy for programmers to create a web application that has various functions such as plugins, and concepts to form a certain system to be arranged and structured neatly. Using a framework does not mean it will be free coding. As a framework user, you are required to use functions and variables that are in a framework that are being used [1]. Python is a multipurpose interpretive programming language with a design philosophy that focuses on code readability. Python is claimed to be a language that combines capability, ability, with a very clear code syntax, and comes with the functionality of a large and comprehensive standard library**.** Application Programming Interface allows developers to integrate two parts of one application with different applications simultaneously, which aims to speed up the development process by providing separate item functions so that developers do not need to create similar features [10].

a) Keyword Sheet Display

Figure 1 is a view of the sheet used to set keywords that will be Scraping. Writing keywords on the sheet titled “keyword”

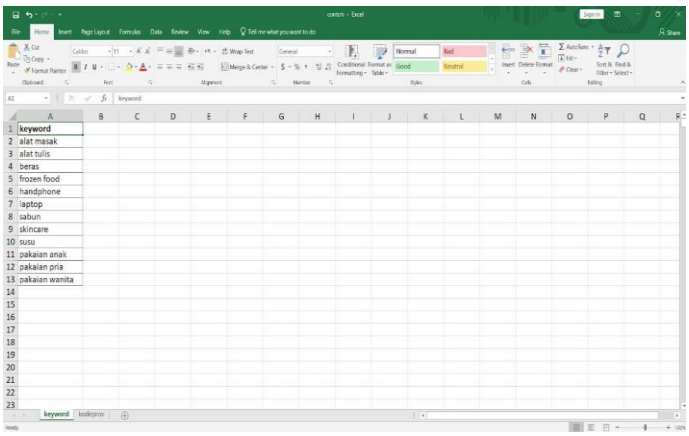

Figure 1. Keyword Sheet Display

Figure 1 is a display that is used to set the location of product sales that will be in scraping, for the location specifically for the west Java Province. There are 2 columns in the “kodeprov” sheet. The first column is the name of the city that will be scraping. Location “code” the district code can be seen in the url displayed when the user clicks on the checklist of the selected district. The location code is a combination of letters and numbers listed after the “cities[]=” parameter.

b) Location Code URL Display & Scraping Process Completed

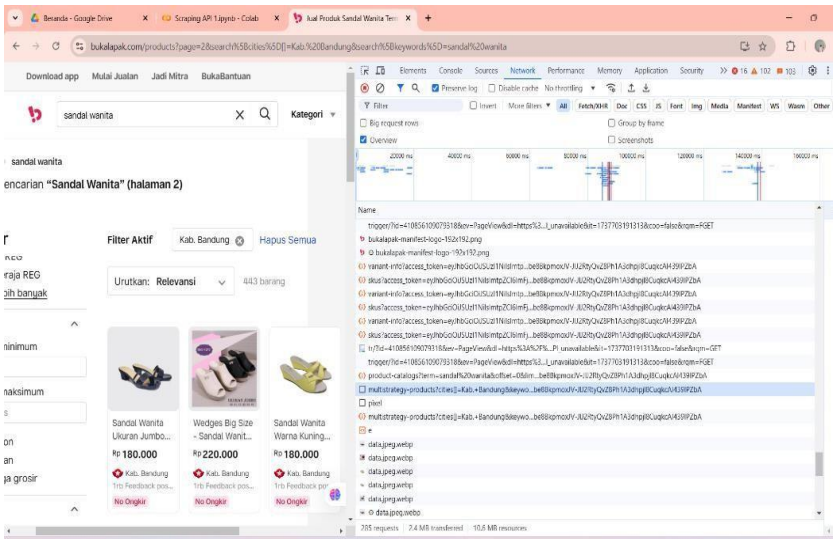

Figure 2. Location Code URL Display

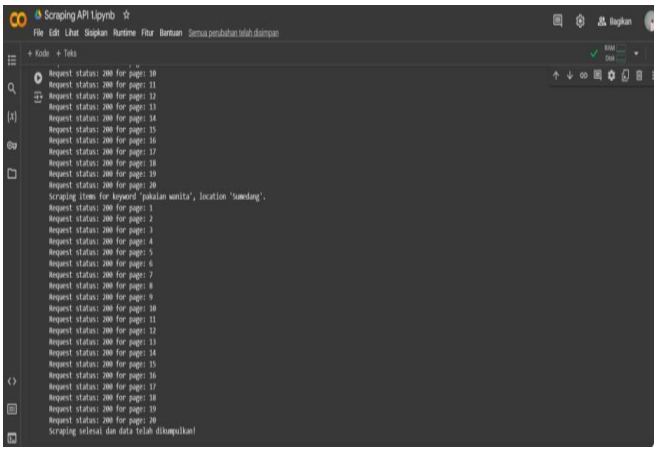

Figure 3. Scraping Process Completed

c) Excel based API scraping results

The results of this scraping API get 74796 data from 12 categories. The following is attachment that are showed on figure 4.

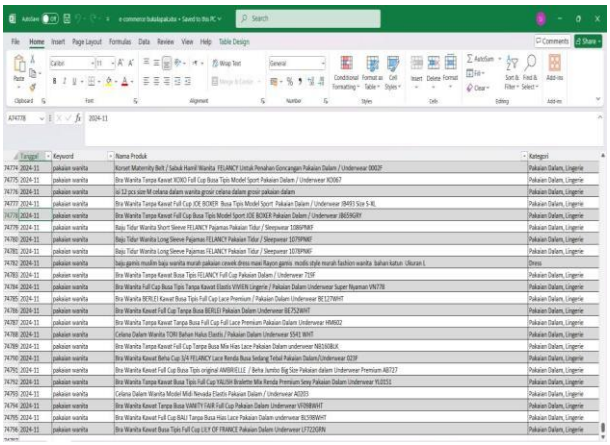

Figure 4. Excel based API scraping results

Scraping data from Bukalapak e-commerce can be used to support visualization of market trends and analysis of community consumption patterns in Wet Java. Based on the code that has been given, the data retrieved includes important attributes of such as product name, category, original price, current price, product URL, location, average rating, and number of user reviews. Once this data is collected through the process of scraping using Bukalapak API, the data can be processed into the visualization platform for further analysis. For example, by utilizing product category attributes, can also group based on certain categories to see the consumption trends of people in each region. Price data can be analyzed to understand the distribution of prices in the West Java region, while location data is used to evaluate the distribution of availability of goods based on geographical areas. For example, data that has been scraped with Python scripts can be used to create bar charts that show the average price product category or the number of customer reviews in each category. The data collection process is run in Google Colab with Google Drive integration for data storage. Google Collab provides the flexibility to manage data in the cloud, allowing visualization directly after data is scraped.

d) Initial price visualization

This histogram diagram shows the initial distribution of products. Most of the left side of the chart. The reflects that the majority of starting prices are in the relatively affordable range. However, there are some products with much higher starting prices, which creates a long tail to the right of the graph. This pattern shows an uneven distribution, where most of the data is concentrated at low prices, while only a few products have premium prices. This form of distribution is known Skewed Right. The diagram shows in figure 5.

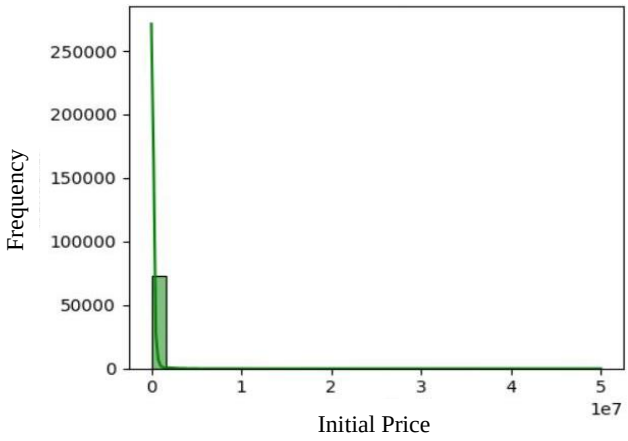

Figure 5. Initial price visualization

e) Current Price Chart Visualization

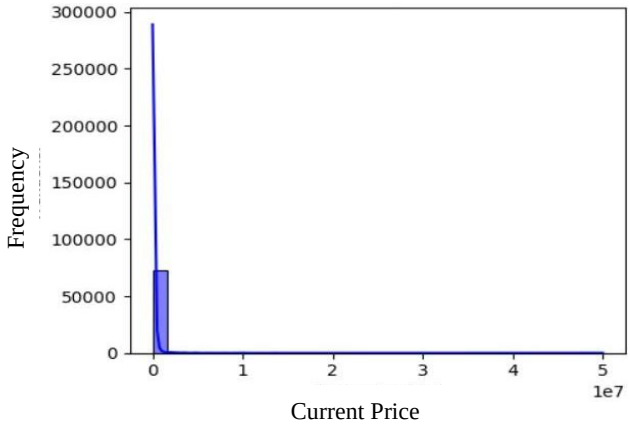

Figure 6. Current Price Chart Visualization

In this figure 6 shows the current price distribution of the products. The majority of products have very low current price distribution of the products. The majority of products have low current prices, as can be seen from the high frequency in the left part of the cart. The high frequency indicates that most of the products are within the affordable price range. However, the graph has a long tail to the left, indicating that there area a small number of products with prices that are much higher than the majority. This distribution, known as Skewed Right, indicates that there are some products with premium prices or unusually high prices, although they are few in number.

f) Visualization of 10 Categories

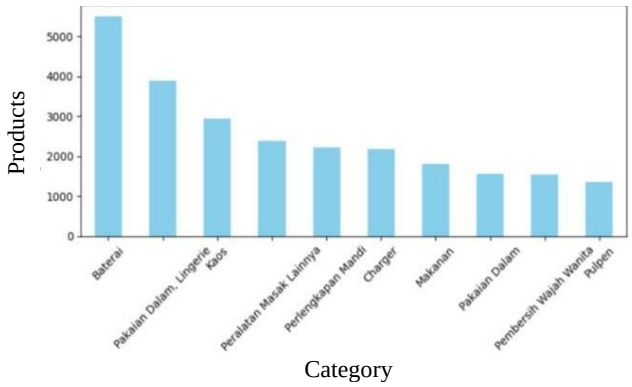

Figure 7. Visualization of Top 10 Product Categories

Figure 7 shows the top 10 product categories based on the number of products available in the Bukalapak marketplace.

g) Top 10 Locations

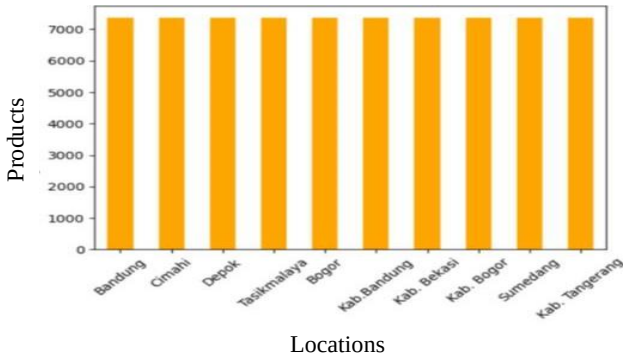

Figure 8. Top 10 Locations based on Products

The diagram shows the 10 locations with the highest number of products. The diagram shows on figure 8. Each location such as Bandung, Cimahi, Depok, and Tangerang. It has almost the same number of products, which is around 7000 products. This indicates that the distribution of products in the marketplace is quite evenly distributed across these cities and regencies. Big cities such as Bandung, Depok and Bogor show high trading activity, possibly due to the large number of sellers and vast market in these areas.

h) Stock Count Visualization

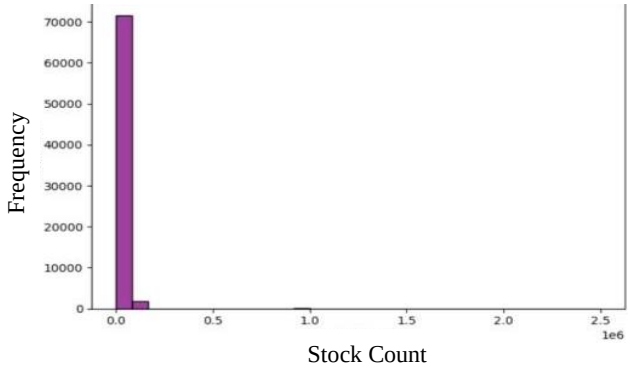

Figure 9. Stock Count Visualization

This diagram shows the distribution of product stock Bukalapak marketplace. Most Product have a very low amount of stock, with a high frequency in the small stock range (near 0). Meanwhile, only a few products have very large stocks (more than 1 million). This pattern is common in marketplaces, where the majority of products are sold in small quantities, such as exclusive or made-on demand products. Products with large stocks are likely to be mass or popular items with high demand. This distribution indicates that product stock is uneven, with the majority of products having limited stock. This could be an indication of the seller's strategy to maintain efficient stock management.

i) Relationship between Number of users and Average Rating

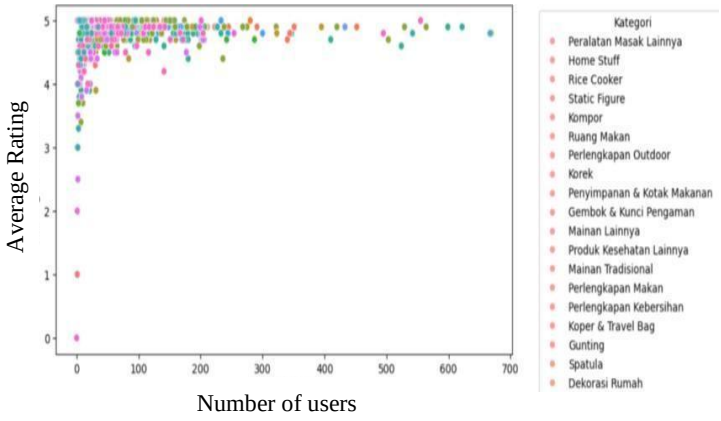

Figure 10. Relationship between Number of users and Average Rating

The diagram above illustrates the relationship between the number of users who gave reviews and the average rating on various product categories. The horizontal axis (X) shows the number of users, while the vertical axis (Y) represents the average rating given, with a scale from 0 to 5. Each colored dot in the diagram represents a product category described in the right-hand side. Based on the data, it can be seen that most categories have a number of users below 200 and an average rating above 4, indicating that the majority of products receive positive ratings. Some categories have a much higher number of users (above 400), but still show a good average rating, in the upper range of the rating scale. There is no linear relationship between the number of users and the average rating, so a large number of reviews does not necessarily affect the average rating of a category.

j) Top 10 Keywords by number of products

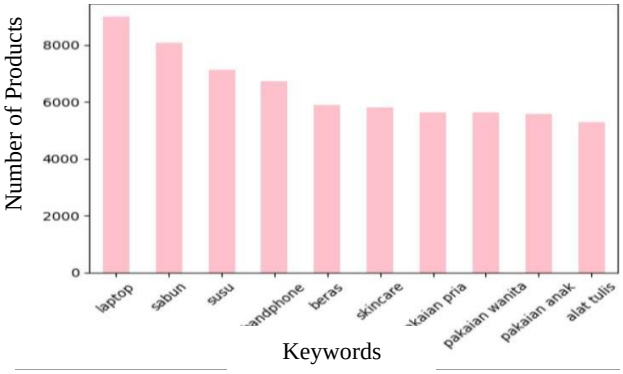

Figure 11. Top 10 Keywords by number of products

This diagram shows the top 10 keywords based on the number of products available. The horizontal axis (x) display keywords, such as “laptop”, “soap”, “milk”, and others, while the vertical axis (Y) represents the number of products listed for each keyword. The keyword “laptop” is in the first position with the highest number of products, which is more than 8,000.

4. CONCLUSION

This research shows that the application of web scraping techniques using API and python-based automation can be effectively used to collect product data from e-commerce sites, especially Bukalapak. The scrapping process run through the Google Collab platform is able to generate 74,796 product data from 12 categories, which include attributes such as price, category, seller location, and customer reviews. The results of data collection are then analyzed and visualized to identify market trends, geographical distribution of products, and consumption patterns of people in West Java Province. This approach not only accelerates the real-time data acquisition process, but also supports efficiency and accuracy in market statistics analysis. The application of this method in agencies such as the Central Bureau of Statistics of West Java Province can be an innovative solution to support data based policy making in the digital economy era.

5. ACKNOWLEDGEMENTS

The author would to thank all those who have supported and contributed to the completion of this research. First, the entire academic community of the informatics Engineering Study Program at UIN Sunan Gunung Djati Bandung and UIN Alaudin Makassar. All of people that support in completing this research.

REFERENCES

[1] S. Kusumo and R. Somya, “Penerapan Web Scraping Deskripsi Produk Menggunakan Selenium Python dan Framework Laravel,” *JATISI (Jurnal Teknik Informatika dan Sistem Informasi)*, vol. 9, no. 4, pp. 3426–3435, Dec. 2022, doi: 10.35957/JATISI.V9I4.2727.

[2] A. Pramesti, C. Novitasari, and D. Oktaviani, “Penerapan Manajemen Operasional di Era Digital dan Perkembangan E-Commerce,” *Economics Business Finance and Entrepreneurship*, pp. 88–97, Aug. 2023, Accessed: Apr. 26, 2025. [Online]. Available: https://proceedings.ums.ac.id/ebfelepma/article/view/3111

[3] S. Fatimah and S. K. Mukarramah, “Pendampingan Pengembangan Pemasaran Digital di Desa Wisata Rinding Allo,” *Jurnal IPMAS*, vol. 3, no. 3, pp. 165–173, Dec. 2023, doi: 10.54065/IPMAS.3.3.2023.360.

[4] Matthew. Turland, “Phparchitect’s guide to web scraping with PHP,” p. 173, 2010.

[5] D. Glez-Peña, A. Lourenço, H. López-Fernández, M. Reboiro-Jato, and F. Fdez-Riverola, “Web scraping technologies in an API world,” *Brief Bioinform*, vol. 15, no. 5, pp. 788–797, Sep. 2014, doi: 10.1093/BIB/BBT026.

[6] A. Z. Rizquina and C. I. Ratnasari, "Implementasi Web Scraping untuk Pengambilan Data Pada Website E-Commerce," *Jurnal Teknologi Dan Sistem Informasi Bisnis*, vol. 5, no. 4, pp. 377–383, Oct. 2023, doi: 10.47233/JTEKSIS.V5I4.913.

[7] R. Gunawan, A. Rahmatulloh, I. Darmawan, and F. Firdaus, "Comparison of Web Scraping Techniques : Regular Expression, HTML DOM and Xpath," pp. 283–287, Mar. 2019, doi: 10.2991/ICOIESE-18.2019.50.

[8] E. S. Sulistiyawati and A. Widayani, "Marketplace Shopee Sebagai Media Promosi Penjualan UMKM di Kota Blitar," *Jurnal Pemasaran Kompetitif*, vol. 4, no. 1, p. 133, Oct. 2020, doi: 10.32493/jpkpk.v4i1.7087.

[9] A. Triono, A. S. Budi, R. Abdillah, and Wahyudi, "Implementasi Peretasan Sandi Vigenere Chipher Menggunakan Bahasa Pemrograman Python," *JOCITIS-Journal Science Infomatica and Robotics*, vol. 1, no. 1, pp. 01–09, Sep. 2023, Accessed: Apr. 26, 2025. [Online]. Available: https://jurnal.ittc.web.id/index.php/jct/article/view/28

[10] I. Irian and Y. Yudhistira, "Implementasi Application Programming Interface (API) Kawal Corona Sebagai Media Informasi Pandemik Covid-19 Berbasis Android: Array," *Jurnal Sistem Informasi dan Teknologi Peradaban*, vol. 2, no. 1, pp. 22–29, Jul. 2021, Accessed: Apr. 26, 2025. [Online]. Available: https://journal.peradaban.ac.id/index.php/jsitp/article/view/755

[11] T. Cut *et al.*, "Implementasi Sistem Informasi Penjualan Produk Elektronik Berbasis Web dengan Menggunakan Laravel Framework," *Buletin Poltanesa*, vol. 20, no. 2, pp. 51–56, Dec. 2019, Accessed: Apr. 26, 2025. [Online]. Available: https://www.neliti.com/publications/338568/

[12] V. Alhafiz and M. A. Adiguna, "Development of Web Scraping Application for E-Commerce Product Price Data Collection Using Python (Case Study: Tokopedia)," *JUPIK : Jurnal Penelitian Ilmu komputer*, vol. 2, no. 2, pp. 1–13, Jun. 2024, Accessed: Apr. 26, 2025. [Online]. Available: http://mypublikasi.com/index.php/JUPIK/article/view/98

[13] Terry. Felke-Morris, "Web development & design foundations with XHTML," p. 651, 2009.

[14] W. B. Zulfikar, M. Irfan, M. Ghufron, Jumadi, and E. Firmansyah, "Marketplace affiliates potential analysis using cosine similarity and vision-based page segmentation," *Bulletin of Electrical Engineering and Informatics*, vol. 9, no. 6, pp. 2492–2498, Dec. 2020, doi: 10.11591/EEI.V9I6.2018.

[15] A. R. Atmadja, M. N. Rahman, and W. B. Zulfikar, "Permodelan Topik pada Layanan Akademik Perguruan Tinggi dengan Menggunakan N-Gram," *INTERNAL (Information System Journal)*, vol. 7, no. 2, pp. 167–177, Dec. 2024, doi: 10.32627/INTERNAL.V7I2.1192.

[16] R. Wiryasaputra, A. Salomo, N. Sevani, and Seruni, "Peningkatan Pola Berfikir Komputasi pada Siswa/i SMAK Mater Dei Melalui Bahasa Pemograman Java dan Python," *Servirisma*, vol. 2, no. 2, pp. 127–145, Nov. 2022, doi: 10.21460/SERVIRISMA.2022.22.28.