International Journal of Trend in Scientific Research and Development (IJTSRD)
Special Issue on Advancements and Emerging Trends in Computer Applications - Innovations, Challenges, and Future Prospects
Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Web Scrapping Based Flipkart Data Analytics and Visualization

Prachi J. Wanjari

PG Student, Department of Computer Application, G. H. Raisoni University, Amravati, Maharashtra, India

ABSTRACT

Large amounts of data are produced by e-commerce companies like Flipkart, which offer valuable insights into consumer attitudes, industry trends, and product performance. Using Beautiful Soup and Scrapy for data crawling, Pandas and NumPy for data pre-processing, and Matplotlib and Seaborn for visualization, the current study focuses on online scraping-based Flipkart data analysis and visualization. Text Blob, Vader, and LSTM are examples of Natural Language Processing (NLP) systems that help with sentiment analysis of customer evaluations. The system is a web application built on Streamlit that provides interactive dashboards for analysing sales trends, sentiment data, and price movements. Future advances would involve machine learning for demand prediction and expanding data collection across several e-commerce platforms to improve market decision-making. The research integrates web scraping, analytics, and business intelligence.

KEYWORDS: *Web Scrapping, Data Analysis, Data Visualization, Data Mining, Flipkart data Extraction, Beautiful Soap, Scrapy, price comparison, E-commerce Analytics, Python Web Scrapping, Consumer.*

I. INTRODUCTION

The accelerated growth of e-commerce has led to huge volumes of web data that can be used for decision-making based on data. Sites such as Flipkart hold product-related data of immense value, such as pricing trends, customer feedback, and sales trends, which, if properly analysed, can provide valuable business insights. Collecting and analysing such huge volumes of data manually is not feasible, so automated data collection through web scraping is the way to go. This work is dedicated to web scraping-based Flipkart analytics and visualization, which seeks to scrape real-time data from Flipkart, analyze the different product attributes, and display insights in an interactive manner.

The key aims of the current research are to design an automated web crawling system for harvesting Flipkart product information, employing data pre-processing mechanisms for cleansing and formatting the collected data, and conducting data analysis for price trends, customer reviews, and demand patterns. Beautiful Soup and Scrapy Python libraries are utilized for web crawling, while Pandas and NumPy help in processing the data. Visualization libraries such as Matplotlib and Seaborn display the insights obtained graphically. Moreover, Natural Language Processing (NLP)-based sentiment analysis is used on customer reviews to determine user satisfaction and product perception. For improving user usability, the system is implemented as a Streamlit web application to provide interactive analysis of Flipkart data. The web application includes search facilities, graphical plots of price fluctuation over time, sentiment reports, and category-based performance analytics. The primary outputs of this research are an automated data pipeline for ongoing extraction, complete analytics dashboard, and sentiment analysis module for interpreting customer reviews. This research fills the gap between web scraping, data analysis, and business intelligence by presenting an AI-powered methodology for analysing e-commerce data. Future extensions include broadening the scope to other e-commerce sites, embedding machine learning algorithms for demand prediction, and tuning real-time analysis for better decision-making.

II. RELATED WORK

The research cites the difficulty of extracting large amounts of web data manually for analysis. It uses a web scraping method with the help of automation tools such as ParseHub, showing how organized data can be effectively gathered. Outcomes indicate enhanced data accessibility, efficiency in automation, and wide applicability across sectors such as e-commerce and research [1]. The research tackles web scraping detection and blocking by websites. It discusses less noticeable web scraping methods, IP rotation, request randomization, and headless browsing for evading detection. Outcomes show improved scraping success, lower blocking occurrences, and more efficient data extraction for continuous web scraping processes [2]. The work recognizes scalability challenges in conventional web scraping for big data applications. It suggests a cloud-based web scraping method, with distributed computing and cloud storage to manage large-scale data scraping. Experiments prove improved scalability, high-speed processing, and enhanced efficiency in the management and analysis of huge amounts of scraped data [3]. The research responds to the challenge of product price and supply trend analysis in darknet markets. It suggests a framework of analysis based on web scraping and data analysis for tracking market dynamics. Results show successful price monitoring, supply fluctuation analysis, and enhanced darknet economic pattern understanding for cybersecurity and law enforcement purposes [4]. The research detects inefficiencies in convenience store distribution because of a lack of real-time data. It suggests using web scraping technique along with Google API services to maximize location-based distribution. Outcomes show enhanced decision-making, supply chain efficiency, and inventory management by means of automatic data gathering and analysis [5]. The issue addressed in this research is the difficulty of scraping and processing large-scale e-commerce data for business intelligence. The article takes a web scraping-based solution with Python, using Beautiful Soup and Scrapy for data extraction and Pandas, NumPy, Matplotlib for processing. Outcomes show effective data extraction, organized analysis, and meaningful visualization, promoting informed decision-making [6]. The research points out the challenge of collecting online data

for psychological studies. It employs a web scraping method with R, utilizing rvest and other R packages for data collection and analysis. Results indicate successful automation of data collection, enhanced research efficiency, and better data-driven insights for psychological research [7].

III. DATA AND SOURCES OF DATA

Flipkart data is gathered for analytics by web scraping with programs like Beautiful Soup, Scrapy, or Selenium; however, legal constraints and permissions must be considered. As an alternative, structured data can be obtained from datasets found on Kaggle or Google Dataset Search. Flipkart's API offers a formal way to get product details, costs, and reviews if it is available. Price patterns, consumer feedback, and top-rated products are revealed through analysis using the Python modules Pandas, Matplotlib, Seaborn, and Plotly. Word clouds, bar charts, and heatmaps are examples of visualizations that enhance insights and make data useful for business decision-making, competitive analysis, and market research.

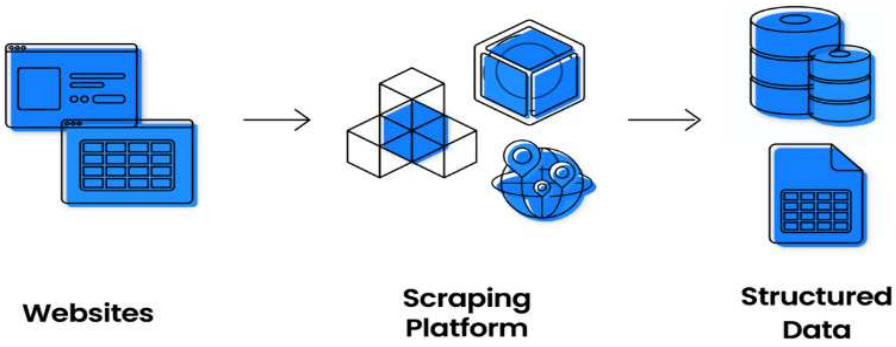

Fig:1 Extracting Data from Websites to Structured Formats

Web scraping is the automated process of extracting data from websites and converting it into structured formats like databases or spreadsheets. It involves fetching web pages, parsing the content, and extracting relevant information using tools like BeautifulSoup, Scrapy, or Selenium. The scraped data is then cleaned and stored for analysis, business intelligence, or machine learning. A scraping platform automates this workflow, ensuring efficient data collection. Structured data from web scraping is valuable for market research, price monitoring, lead generation, and more. However, ethical considerations and website terms of service should always be followed when scraping.

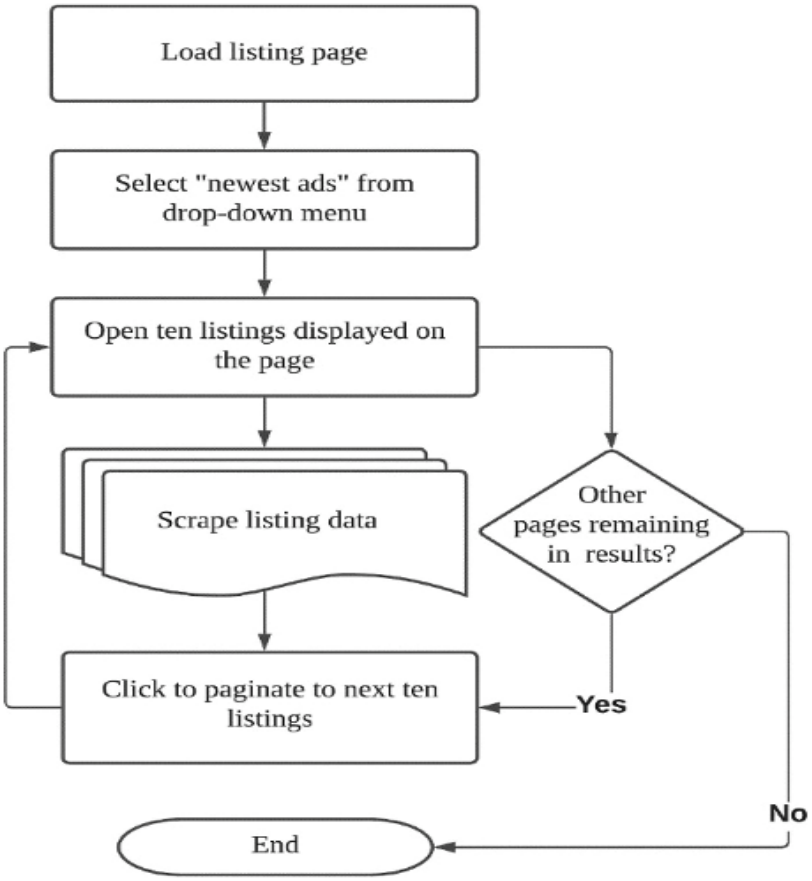

Fig: 2Web Scraping Flowchart for Listing Data Extraction

IV. RESEARCH METHODOLOGY

The process starts with the loading of the listings page, whereby the scraper picks the "newest ads" choice from a drop-down menu to ensure that only the newest listings are obtained. It then goes on to retrieve ten listings shown on the page and obtains

relevant information such as price, location, and property details. Once data collection is done, the system checks if there are other pages within the search results. When there are additional pages, the scraper advances to the next batch of ten listings and extracts again. This process is repeated until there are no pages left to be processed, when the operation terminates. The flowchart illustrated below represents a structured and automated approach to web scraping, allowing for effective data gathering without much manual intervention. Through systematic going through of the listings, the system makes possible comprehensive data extraction to be used for market research, trend analysis, and decision making in the real estate sector.

Equations

1. Missing Value Percentage

$$\text{Missing data\%} = \frac{\text{Total Missing Entries}}{\text{Total Entries}} X100$$

(Used to calculate missing values in the dataset.)

2. Data Cleaning Efficiency

$$\text{Cleaning Efficiency} = \frac{\textit{Total Errors Before Cleaning} \quad \textit{Total Errors After Cleaning}}{\textit{Total Errors Before Cleaning}} X100$$

(Used to measure effectiveness of data cleaning.)

3. Data Normalization (Min-Max Scaling)

$$X` - \frac{X - X_{min}}{X_{max} - X_{min}}$$

(Used to normalize product attributes like price and ratings.)

4. Data Standardization (Z-Score**)**

$$Z = \frac{X - \mu}{\sigma}$$

(used to standardize numerical values like product prices.)

5. Sentiment Polarity Score (TextBlob)

$$\text{Polarity Score} = \frac{\sum(\textit{Positive Words}) - \sum(\textit{Negative Words})}{\textit{Total words in Review}}$$

(Help determine the positivity or negativity of a review.)

6. Vader Sentiment Score

Compound Score=Positive Score=Negative Score

(Used for sentiment classification.)

V. RESULT & DISCUSSION

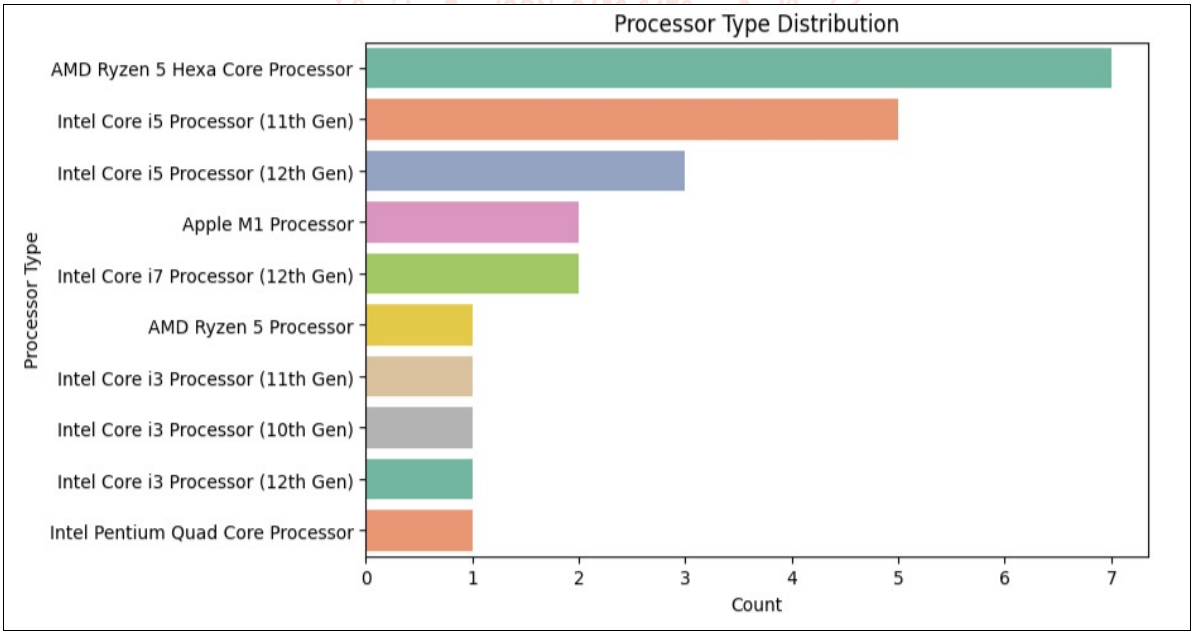

Fig:3 Distribution of Processor Types in Laptops

This bar chart shows the frequency of various processor types acquired using web scraping. AMD Ryzen 5 Hexa Core Processor occurs the most, then Intel Core i5 (11th Gen), and Intel Core i5 (12th Gen). The remaining processors such as Apple M1, Intel Core i7 (12th Gen), and Intel Core i3 models occur but in smaller numbers. The information was probably gathered from websites selling products, product pages, or technology review websites to identify trends in processor popularity. It can be helpful for market intelligence, consumer preference information, or suggesting products based on recent availability and demand.

Find the product having highest rating?

```
In [61]: Laptop_Flipkart[Laptop_Flipkart['Ratings']==max(Laptop_Flipkart['Ratings'])]
```

Out[61]:

	ProductName	Stars	Ratings	Reviews	CurrentPrice	MRP	Processor	RAM	Windows	Storage	ImageURL
15	Apple 2020 Macbook Air Apple M1	4.7	13063	1087	₹79990	₹99900	Apple M1 Processor	8 GB DDR4 RAM	Mac OS Operating System	256 GB SSD	https://rukminim2.flixcart.com/image/312/312/k...
18	Apple 2020 Macbook Air Apple M1	4.7	13063	1087	₹79990	₹99900	Apple M1 Processor	8 GB DDR4 RAM	Mac OS Operating System	256 GB SSD	https://rukminim2.flixcart.com/image/312/312/k...

Find the product having highest review?

```
In [62]: Laptop_Flipkart[Laptop_Flipkart['Reviews']==max(Laptop_Flipkart['Reviews'])]
```

Out[62]:

	ProductName	Stars	Ratings	Reviews	CurrentPrice	MRP	Processor	RAM	Windows	Storage	ImageURL
15	Apple 2020 Macbook Air Apple M1	4.7	13063	1087	₹79990	₹99900	Apple M1 Processor	8 GB DDR4 RAM	Mac OS Operating System	256 GB SSD	https://rukminim2.flixcart.com/image/312/312/k...
18	Apple 2020 Macbook Air Apple M1	4.7	13063	1087	₹79990	₹99900	Apple M1 Processor	8 GB DDR4 RAM	Mac OS Operating System	256 GB SSD	https://rukminim2.flixcart.com/image/312/312/k...

Fig4: Identification of the Laptop with the Highest Rating and Reviews

The review determines the Apple 2020 MacBook Air (M1 Processor) to be the highest-rated laptop with the most reviews. With a 4.7-star rating from 13,063 ratings and 1,087 reviews, the product has witnessed strong customer satisfaction and extensive use.

The excellent rating (4.7 out of 5) indicates that customers have had a good experience with this laptop, most probably as a result of performance, battery life, and overall user experience with the Apple M1 chip. The large volume of reviews (1,087) also indicates high customer interest, demonstrating that numerous customers were eager to share their experience.

The findings reveal that Apple's MacBook Air (M1) is a top pick among customers, hence a reference point for comparison with other laptops in performance, price, and customer satisfaction in the market.

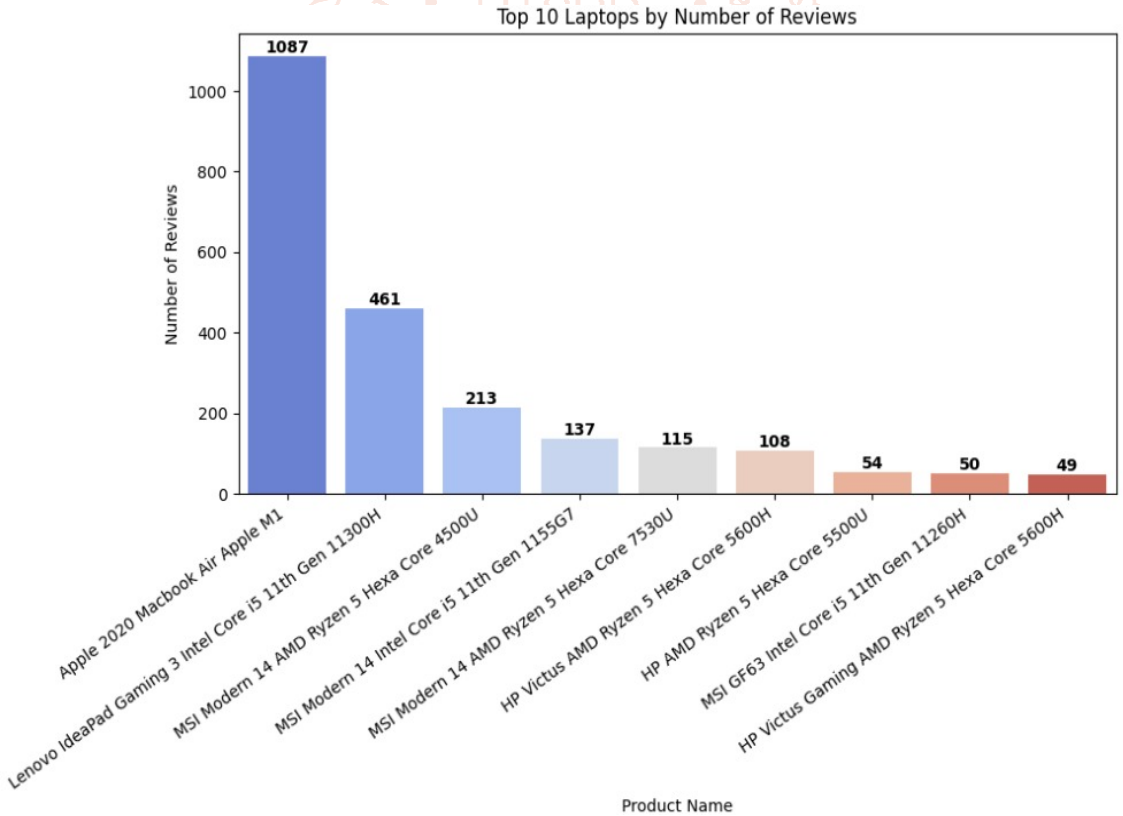

Fig:4 bar chart result based on customer review

Analysis of Laptop Popularity Based on User Reviews

Comparing the reviews of the top laptops, one product stands far above the rest – Apple 2020 MacBook Air with M1 chip. At 1,087 reviews, this is the lowest product in the list, meaning customers are extremely passionate about it, which could be attributed to the brand, efficiency or ecosystem, and this is evidenced in the examination of the comments. Second position is held by the Lenovo IdeaPad Gaming 3 with Intel Core i5 11th Gen 11300H. It shows evident inclination towards mid-range gaming laptops particularly among price sensitive consumers. Then it seems that the MSI Modern 14 model is pretty popular, the laptop with AMD Ryzen 5 Hexa Core 4500U has 213 reviews and the laptop with Intel Core i5 11th Gen 1155G7

has 137 reviews. These numbers indicate that there is a good market for thin and productivity-oriented laptops. But where the gaming laptops are concerned, the reviews are not many. HP Victus AMD Ryzen 5 5600H processor has 49 reviews and MSI GF63 Intel Core i5 11th Gen 11260H has 50 reviews. It implies that although gaming laptops are well-liked by some segment of customers, they could not receive much support as general ultraportables or budget laptops. Overall, the statistics show that customers are most active for ultraportable and budget laptops, whereas gaming laptops with high performance are niche.

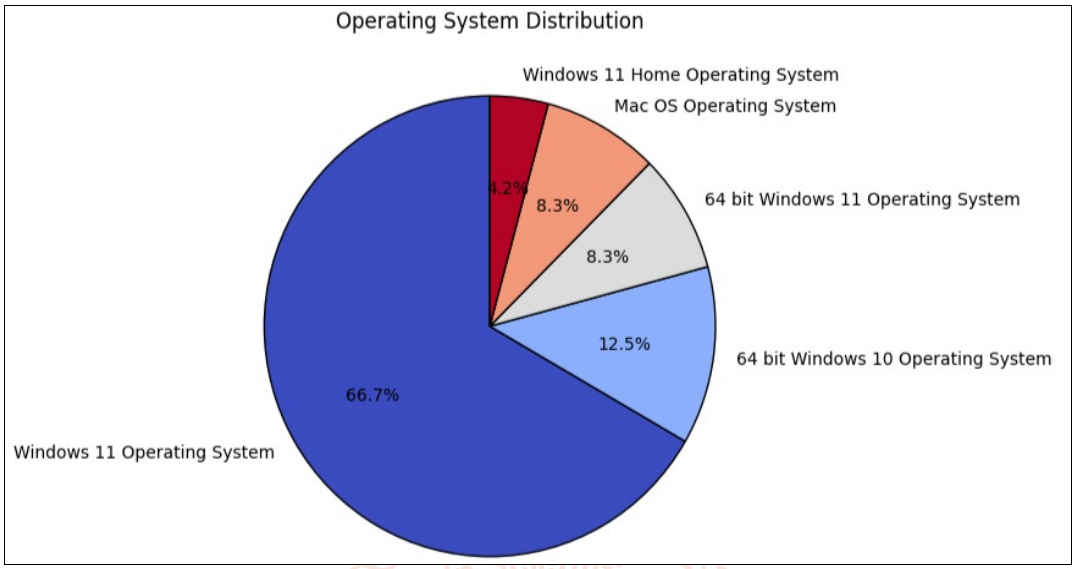

Fig: 5 Operating System Distribution among Laptops

The pie graph illustrates the frequency of operating systems in the evaluated laptops. Results show that the Windows 11 Operating System takes the lion's share of the market at 66.7%. This explains that most customers prefer the most recent Windows platform, perhaps influenced by its tighter security measures, enhanced user experience, and upgraded performance tuning.

Other significant distributions include 64-bit Windows 10 (12.5%), which remains in favour among users who might prefer compatibility and stability with legacy applications. Mac OS and Windows 11 Home Edition are also each represented at 8.3%, which reflects a niche but relevant preference for Apple devices and the light version of Windows 11. The least significant share (4.2%) is from other versions of Windows 11, reflecting minimal use of these alternatives.

The findings emphasize that Windows 11 is now the default option for laptop users, indicating its popularity in recent iterations. Nevertheless, the availability of Mac OS indicates a loyal clientele for Apple products.

VI. CONCLUSION

This project is able to showcase the capability of web scraping, data analysis, and visualization in extracting and analysing Flipkart laptop data. We automated data extraction using Python libraries such as BeautifulSoup and Scrapy, data pre-processing and analysis were done using Pandas, and Matplotlib and Seaborn were employed for visualizing major trends in processor types, operating systems, ratings, and prices.

Our results indicate that AMD Ryzen and Intel Core processors are most prominent in the market, followed by Windows 11 as the most sought-after operating system. Apple M1 laptops are also picking up pace. The research findings can provide insights to optimize product offerings and price strategies.

Moreover, an interactive Streamlit web application was also created to increase data exploration simplicity. Future extensions involve the aggregation of data across various e-commerce platforms and applying machine learning in demand forecasting. This project reinforces the need for data-driven decisions in comprehending market trends as well as increasing consumer awareness within the e-commerce sector.

VII. REFERENCE

[1] Usha Kosarkar, Gopal Sakarkar, Shilpa Gedam, "An Analytical Perspective on Various Deep Learning Techniques for Deepfake Detection", *1st International Conference on Artificial Intelligence and Big Data Analytics (ICAIBDA),* 10th & 11th June 2022, 2456-3463, Volume 7, PP. 25-30.

[2] Usha Kosarkar, Gopal Sakarkar, Shilpa Gedam, "Revealing and Classification of Deepfakes Videos Images using a Customize Convolution Neural Network Model", *International Conference on Machine Learning and Data Engineering (ICMLDE)*, 7th & 8th September 2022, 2636-2652, PP. 2636-2652.

[3] F. Färholt, "Less Detectable Web Scraping Techniques," Bachelor Thesis, Linnaeus University, Faculty of Technology, Department of computer science and media technology (CM), 2021.

[4] R. S. Chaulagain, S. Pandey, .R. Basnet, and S. Shakya, "Cloud Based Web Scraping for Big Data Applications," 2017 IEEE International Conference on Smart Cloud (Smart Cloud), pp. 138–143, Nov. 2017, doi: 10.1109/smartcloud.2017.28.

[5] Y. Yannikos, J. Heeger, and M. Brockmeyer, "An Analysis Framework for Product Prices and Supplies in Darknet Marketplaces," Proceedings of the 14th International Conference on Availability, Reliability

and Security, Aug. 2019, doi:10.1145/3339252.3341485.

[6] Q. T. Le and D. Pishva, "Application of web scraping and Google API service to optimize convenience stores' distribution," 2015 17th International Conference on Advanced Communication Technology (ICACT), pp. 478–482, Aug. 2015.

[7] D. M. Thomas and S. Mathur, "Data Analysis by Web Scraping using Python," 2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2019, pp. 450-454, doi: 10.1109/ICECA.2019.8822022.

[8] A. Bradley and R. J. James, "Web scraping using R," Advances in Methods and Practices in Psychological Science, vol. 2, no. 3, pp. 264–270, 2019.

[9] R. Gunawan, A. Rahmatulloh, I. Darmawan, and F. Firdaus, "Comparison of web scraping techniques: Regular expression, HTML Dom and xpath," Proceedings of the 2018 International Conference on Industrial Enterprise and System Engineering (IcoIESE 2018), 2019.

[10] L. Richardson, "BeautifulSoup," Crummy, 2020. https://www.crummy.com/software/BeautifulSoup/

[11] Scrapfly, "Web Scraping with Python and BeautifulSoup," ScrapFly, Jan. 03, 2022.