A Demystified Overview of Data Scraping

Mustapha Shehu[1], Mustafa Man [2], Wan Aezwani Wan Abu Bakar [3], Mohd Kamir Yusof [4], and Ily Amalina Ahmad Sabri [5]
[1, 2, 5] Faculty of Ocean Engineering and Informatics, Universiti Malaysia Terengganu, 21030 Kuala Nerus, Terengganu, Malaysia.
[3,4]Universiti Sultan Zainal Abidin, Malaysia
[1]p5265@pps.umt.edu.my, [2]mustafaman@umt.edu.my, [3]wanaezwani@unisza.edu.my , [4]mohdkamir@unisza.edu.my, [5]ilylina@umt.edu.my
Corresponding author email: mustafaman@umt.edu.my

***Abstract*— Data scraping is a concept that involves the extraction of relevant data from a pool of information stored in a computer. Data scraping is universally known as web scraping because the web contains massive amount of information that is easily accessible and extracted. Web scraping is valuable to all field of human endeavour. The paper gives a vivid conceptualization of data scraping and some minor misconception between data mining, web crawling and web scraping. Furthermore, the phases and procedure of data scraping were outlined. The merit of web scraping over API scraping were explicated. Moreover, the numerous software and tools that support the scraping of websites were stated. Even though web scraping has vast prominence, there are also some technical issues and challenges associated with it. Finally, some of the legal and ethical issues related to information extraction were discussed and it is obvious that data scraping is permitted as long as users comply to the terms and conditions of the target site.**

***Keywords*— Data Scraping; Web Scraping; Web crawling; Data mining; API Scraping.**

1. Introduction

Data Scraping or web scraping is a fundamental and indispensable process used in the collection of pertinent and up-to-date information from websites. This procedure of information extraction is progressively getting more and more attention due to its prominence in the 21st century. Data scraping enhances qualitative and timely decision making.

As of 31st July 2022, there were more than 5 billion people around the globe using the internet (Internet World Stats, 2022). Thus, in every millisecond, an inconceivable amount of information is produced from the utilization of the internet. This perpetual and immense amount of information on the internet is of great significance to different field of human endeavor. Myriad of disciplines and organizations use data scraping technique to extract and utilize relevant information from the enormous amount of data available on the web [1].

This pattern of data gathering help reduce cost and eliminate physical contact. Data scraping involves the collection and conversion of semi- structured or unstructured data into structured data for further analysis and decision making.

2. Concept of Data Scraping

The importance of conceptualization in every field of study cannot be overemphasized. Terminologies are defined to avoid misconception and enhance comprehension. The following are concept of Data Scraping.

Site Scraping, *Web harvesting* and *Web data extraction* are used interchangeably to refer to the collection of unstructured text from social media and other Web sites [2]. Web scraping according to [3,4] is the act of extracting information from a website and storing it on a computer for later analysis. In another definition by [3], Web scraping also involves the use of application programming interfaces (APIs) and interrelated methods to extract data from digital networks or websites and mobile applications. Web scraping, web extraction or web harvesting, is a procedure used to collect data from the Web for onward storage in a file system or database for later retrieval or analysis [5,6]. Furthermore, Web scraping, screen scraping, or information scraping is a description of how web documents are automaticallyaccessed and specific and predefined information are extracted and transformed into a structured format to be stored [7].

Also, an activity that involves automatic or manual retrieval of vital information from numerous web site with or without the consent of the sites owners is known as web scraping [8]. Web Scraping involves the use of computer programs to automatically extract relevant data from the internet for specific purpose [9]. The method of accessing and retrieving an unstructured information from the web, to create an organized information that will be stored in focal database to be dissected in spread sheets is also known as Web scraping or *web scratching* [4].

Web scraping which is also known as data mining according to [10], is the process of aggregating considerable amounts of data from the web and then storing it unto a database for future analysis and subsequent use. The automatic collection and processing of information from the internet can be referred to as web scraping [11]. Additionally, Web scraping is a technique where unstructured data from the web is converted to structured form and then saved for the purpose of analysis in a central spread sheets or database [12]. The process of collecting an unstructured data from social media and other websites is known as site scraping, web harvesting or web data extraction [2].

Succinctly, Web scraping is the process of automatically extracting information from websites [13]. Comprehensively, it is a technique of creating a computer program that will automatically download data from the web, parse, and organize the data in a structured format [14]. Similarly, Web Scraping, Screen Scraping, Web Data Extraction, or Web Harvesting is a technique employed to extract [15]. Finally, according to [16], Web scraping is synonymous with web crawling, data scraping, data mining, and text mining, and it refers to the activity of

extracting data from websites and converting them to a simpler and more malleable format for analysis.

2.1. Web Crawling

Web crawling and Web scraping are the most prominent and common ways used to collect information from the internet. These terms are sometimes used interchangeably, as was use by [16]. Web crawling is the process of accessing web content and indexing it via hyperlinks, thus, only the URL is extracted [7,18]. In addition, Web crawling is the process of using tools (computer programme) to read, copy and store the content of websites for archiving or indexing purposes. Basically, search engines such as Google, Yahoo, bing etc. use crawling tools to look through websites to discover what content they contain and build entries for search engine index. Therefore, Search engines, crawl the web, analyse the online content and compile all the links they find to match the search request [7] and the full content is made available through the hyperlink. Web Crawlers or spiders can also be used as price comparison tools. For instance, Shopbots are programmes that crawl websites to obtain price information from several sellers for the purpose of competitive advantage [7]. Some of the commonly used tools in web crawling include apache nutche, storm crawler, screaming frog, semrush and deep crawl.

Obviously, Web Scraping and Web Crawling are similar as both refer to the collection of information from the web but differ in the type and amount of information they extract.

2.2. Process of Web Scraping

Web scraping can either be manual copy and paste of website contents or automatic extraction of data using computer programmes. The automatic extraction of web content is usually executed by special programs called scrapers. Scrapers are written scripts that accesses websites through list of URLs and then loads the html codes of these websites. Consequently, relevant and predefined data is then extracted and downloaded as a structured data set. [7].

In principle, the scraper emulates a web user by navigating through the sites and extracting the pre-defined information. Thus, keeping the websites traffic at moderate rate, through some delay between the downloads and http request [7].

According to [18], Data scraping process can also be demystified into fetching, extraction and transformation phase as elucidated in figure 1.

In the **fetching phase** the desired web site that contains the relevant information or data is accessed through HTTP protocol. This protocol is used to send a GET requests to the server and receive a corresponding response. Libraries such as cURL or wget can be used in this phase to send an HTTP GET request to the desired URL, thereby getting the HTML document sent back in the response. This is the same methods used by browsers to access web page contents.

The **extraction phase** is the point where the data of interest is extracted from the HTML document that has being fetched. The technologies used in this phase are regular expressions such as HTML parsing libraries or XPath (XML Path) queries which are used to find information in a document.

Finally, the data of interest that has been extracted can now be **transformed** into a structured version, either for storage or presentation. Figure 1 shows the various steps of data extraction phases.

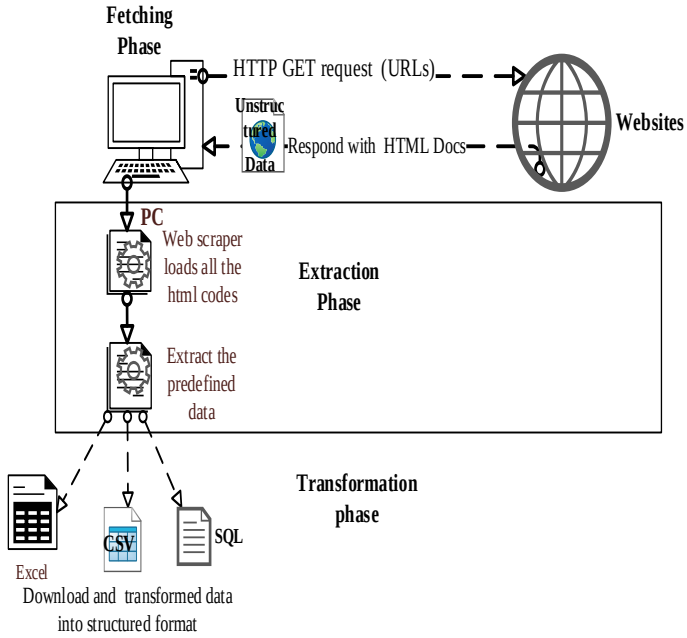

Figure 1 Phases of Data Scraping

Based on the literature reviewed so far, most studies in Data Scraping have only focused on textual explanation of the phases of data scraping. However, few studies such as [19, 20, and 21] use diagram to explain the concept without much elaboration. Thus, the diagram in figure 1 is an attempt to further simplify the three phases of Data scraping process to aid vivid comprehension of the concept.

2.3. Web Scraping Versus API Scraping

Web scraping and API scraping are similar because both perform the function of data extraction. However, they are different in terms of how the extraction is done. Web scraping targets specific HTML or CSS tags to collect data, while API scraping is used to access already available data, that is often delivered in JavaScript Object Notation (JS ON) [3]. Certain websites or web services occasionally offer APIs to fetch or interact with their data [18].

APIs can be used across many sites to easily provide access to information, but this must be at the discretion of the site owner, and they may choose not to make their information accessible through APIs [12]. Although, some website offers API, certainly not all API are free, and in addition, site owners provide restriction in terms of the amount and type of data they provide [12]. Therefore, accessing dataset through an API can either be offered for free or be available at a price and the owner can also limit the number of requests that a single user can make or the amount of data they can access [19].

Web scraping freely extract data from most website through web scraping tools, whereas APIs provide direct access to the type of data that is requested [19]. In a situation whereby Web

scraping is not allowed, as in the case of Facebook, then fetching data through API might be the only option [20]. Web scraping is far faster and more effective than API and can be used to collect and compile data from several websites [12]. In addition, Web scraper can rely on proxy servers [19].

The web scraping tool conveniently organizes the extracted data into a structured format. On the contrary, data obtained through API need to be organized separately. Web scraping is customizable, complex and has the advantage of automatically and repeatedly downloading of a specific set of data [19]. Whereas data obtained from websites through API cannot be updated in near real-time, making the data obsolete [20]. Businesses that require up-to-date information need to use Web scraping technique to acquire data.

Furthermore, a user can remain anonymous when extracting data through Web scraping, however when using API a user needs to register to receive a key and pass it along every time data is requested [20]. Crawling through an unstructured API can be time consuming because one has to deal with queries before getting to the actual data. Although, current websites want to be XHTML validated for the purpose of rankings on search engines, hence the structure is easy to scrape [20]. Obviously, web scraping has more advantages over APIs, however, in the modern era, companies prefer to use both due to their peculiar advantage to extract considerable amount of data. A summary of the differences between Web scraping and API is presented in table 1. These differences are extracted from various authors and compiled in a tabular form to give an overview of the basic differences that exit between the two concepts.

Table 1: Web Scraping Versus API Scraping

Web Scraping	**API Scraping**
Targets specific HTML or CSS tags to collect data	Access already available data, often in JSON.
Data is extracted without offer.	Some Websites offer APIs to fetch or interact with their data.
Any information can be extracted across virtually all sites of interest, without owners' discretion	Only certain information is allowed to be accessed across many sites and in most cases owners' discretion is required.
Data extraction is usually free.	Data extraction is not usually free.
Absence of restrictions.	Site owners provide restriction in terms of the amount and type of data accessed
Web scraping is far faster and more effective	Moderately fast
Can rely on proxy servers	Cannot rely on proxy servers
Web scraping tools conveniently organizes the extracted data into a structured format	Data obtained with the support of API need to be organized separately.
Automatically and repeatedly downloads specific set of data	Not feasible with API
Data can be updated in near real-time.	Data obtained cannot be updated in near real-time
A user can remain anonymous when extracting data.	A user needs to register to receive a key and pass it along every time data is requested.
It is faster to crawl through unstructured data	At times, it is time consuming to Crawl through an unstructured API.

2.4. Software Used in Web Scraping

There are several numbers of software and tools that support the scraping of websites [5]. A web scraping tool is a technology solution that extract data from web sites in a fast, efficient, automated, and structured format [18]. Scrapers can be built using programming languages such as PHP, Node JS, and Python, R etc. Although, developer prefer Python due to its flexibility [8]. There are tools that are readily available for those that lacks programming skills. These tools are integrated with browsers and assist user to navigate to the desired web page and simply point-and-click on required data from the interface. This process circumvents the use of XPath queries or regular expressions for picking out data. There are many scraping tools available online and some are free. Below is a collection of some scraping tools.

Scrapy: is an open-source tool written in Python and runs on Windows, Linux, and Mac. It is originally designed for web scraping but can also be used for web crawling framework and run from command line [5].

Import.io is an online instrument for extracting information from a site without composing code, the client just needs to enter the required URL and the application naturally gets the information that is then stored on Import.io cloud server and further downloaded in CSV or JSON format [4].

Other web scraping tools are Scraping bee, Bright data, Apify, Scrapeowl, Webz.io, Parsehub, Fminer, Data Miner, Chrome extension instant data scraper, Agenty, Selenium, Jaunt, Jsoup, HtmlUnit, PhantomJS, Puppeteer, BeautifulSoup etc.

2.5. Importance of Web Scraping Technique

Data Scraping or Web scraping is a very efficient technique for extracting data from different websites and subsequently use for various purposes, such as web mining, data mining, weather data monitoring etc. Moreover, Web scraping technique is inexpensive and easy to implement. It also has low maintenance cost, higher speed, and accuracy. For instance, Smart chef application use data scraping technique to extract ingredients from various recipe websites. [21].

The cost of production and prices of products can be ascertained through web scraping technologies by accessing stored data. Also, companies can advertise their products to different users based on their data, such as location and cookies settings. Web scraping is currently used for scientific and commercial applications. It aids in the development of big and customized data sets at minimal costs [12].

In the area of food research, web scraping is becoming a promising new method of data collection for food prices. It also

helps to overcome some issues encountered in the traditional data sources, such as official statistics and scanner data [7].

Web scraping is used in many fields as well as for several purposes. For instance, it is commonly used in online price comparisons where data is extracted from several retailers. For example, Shopzilla is a half-billion-dollar company that gathers pricing information from many retailers through scraping and provide users with the opportunity to choose from alternative and best prices [18].

In the Netherlands Web scraping has also been used for official statistics. In one example the Dutch property market was examined for a period of two years, where five different housing websites were monitored with about 250,000 observations per week. This data obtained was later used in market statistics [18].

Furthermore, Web scraping was employed to monitor ten different clothing websites and around 500,000 price observations were collected each day and was used in the production of Dutch Consumer price index (CPI) [18]. Web scraping has also been used in the stock market sector to present a visualization of price fluctuations over time. In addition, social media feeds have been scraped to investigate public opinions [18].

Web scraping is used to search for grey literatures. For example, many academic databases that hold information on research papers may lack some specific details, thus web scraping is used to breach that gap by collecting that extra information elsewhere [22].

In the field of biomedical research, accessing web resources is very important. APIs and or Web scraping are the standard way in which data are collected from these biomedical web resources depending on the scenario. There are scenarios where web scraping is more beneficial, especially if the published APIs deny the use of a desired tool to access required data [18].

Also, in psychological research, web scraping technology is used to gather data from a public depression discussion board (Healthboards.com). This is done to examine the correlations between gender and the number of received responses [23].

In the area Technical and Professional Communication (TPC) content areas, web scraping has a wide area of applications such as gathering social media data, content metadata, user reviews, tutorial feedback, online comments, and discussion forum responses [3].

Web scraping is also used to provide clean news article which is extracted from HTML page from three news websites [24].

Web scraping is also used in the textiles industry. For instance, 68 text-based fields that describe 24,701 clothes were extracted from a major online retailer website. These data were extracted at a rate of approximately 1000 piece of clothing per hour and managed through a Kibana Elasticsearch interface [25].

The application of Web Scraping in every field of human endeavour cannot be over emphasize. It is fast becoming a means of collecting and utilizing unstructured and redundant data for analysis.

2.6. Challenges of Web Scraping

Scrapers are very important extraction tool that support research and analysis. However, the development and deployment of such programs requires a high level of proficiency in programming languages [11]. Although Scrapers are readily available at a cost or free, the quality of information extracted in terms of validity, reliability and interpretation is another area of concern especially in the field of research [11].

In addition, the activities of scraping content from the web tends to overload the target site with high demand or request which could temporarily crash the specific site [8]. The extraction of information at a very fast rate with too many server requests could initiate a denial of service (DoS) attack by the target website. [3]. For instance, websites like Amazon blocks IP address with request that exceed certain threshold.

Moreover, scraping a website often generates a huge amount of data that requires enough storage facility such as Data Warehouse. Thus, a poorly built Data warehouse infrastructure can lead to delay in querying, searching, filtering, and exportation of data. Companies that are engaged in large-scale data extraction tend to overlook this very crucial aspect.

Another contending issue of Web Scraping is the occasional upgrade of User Interface by website owners to improve user experience. Any upgrade by websites without a corresponding adjustment by the Scraper could lead to incomplete data extraction or crashing of the scraper. This is because Web scrapers are developed in line with HTML and JavaScript code elements present in websites [3]. Moreover, some client-side technology such as Ajax and JavaScript that is used for dynamic content generation makes web scraping a difficult task [3].

Furthermore, Anti-Scraping technologies are used by some websites to aggressively prevent any scraping attempt. LinkIn is an example of such sites that deploy ant-scraping technology. Consequently, it cost time and money to develop a technical solution to overcome such anti-scraping technology [3].

Completely Automated Public Turing test (CAPTCHA) is another challenge faced by web scraping. Most E-commerce companies and other websites use CAPTCHA to identify and block internet robots from gaining access to it website. CAPTCHA is one of the most complex challenges faced by web scrapers. Developers seldom overcome this obstacle but often slows down the scraping process [3].

2.7. Ethical and legal Issues of Web Scraping

There are plethora of tools and technologies that are used for Web Scraping purposes. Although, the legality and ethics of data extraction from the Web are still relatively a grey area. Although, existing legal frameworks is to a certain extent applied to Web Scraping. However, the ethical issues surrounding the Web Scraping have largely been ignored [26].

Currently there is no legislation that directly addresses Web Scraping. However, it is guided by a set of related, fundamental legal theories and laws, such as “copyright infringement”, “breach of contract”, the Computer Fraud and Abuse Act (CFAA), and “trespass to chattels [26]. Hence, data gathering or collection through web scraping could have some potential legal implication. For instance, some organizations that placed

their data online may explicitly prohibit web scraping of their site or deny access to robots and other types of automatic data harvesting through 'terms of use' [11,[26]. Consequently, failure to abide with these terms of use may lead to a "breach of contract" on the side of the website's user [26]. Several websites currently require users to agree to "terms and conditions" that explicitly prohibit data scraping of their sites [27]. Moreover, republishing a Scraped data that is owned and explicitly copyrighted by the website owner can lead to a "copyright infringement" [26].

Another legal issue that can arise in the context of web scraping is personal privacy. The extraction of personal data websites may potentially result in the violation of personal privacy rights - even if that information is publicly available [11].

Furthermore, scraping a website at enormous scale could mean more request per second which could harm the targeted site. In a court of law this action could be translated as DDoS attack and the person responsible for the damage can be prosecuted under the "trespass to chattels" law in the US [20]. It is a prevalent perception that, any website that appears on google can be scrapped. On the contrary, it is not advisable to extract information from websites whose robot.txt explicitly warns against it.

Finally, based on the above assertion, web scraping falls within the bounds of legality, if the site's terms of service are obeyed, classified information is not scraped, and the site is not overburdened [20].

3. Discussion

Data Scraping or Web scraping is also known by other names such as Web harvesting, Web data extraction, Site Scraping, Screen Scraping, information scraping and Web scratching. Intriguingly, the various names of web scraping points towards the central meaning of web scraping, which is, the automatic extraction of relevant information from the web. Albeit there is a slight difference between data scraping and web scraping. The former is more versatile, it supports the extraction of information from both online and offline sources, while the latter is specific to the extraction of information from website.

One common misconception is the use of Web scraping and data mining interchangeably as was reported by [10,16]. Conversely, data mining involves advanced pattern-matching algorithms and high-end statistical analyses to assist in identifying trends and patterns [18]. Thus, data mining focuses on data analysis [12]. Another misconception is the use of web scraping and web crawling interchangeably. This misconcepti on could be attributed to the fact that web scraping and web crawling both extract information from the web. However, they vary in the type of information they extract, and tools used.

Moreover, web scraping progresses through distinct phases. The first phase which is the fetching phase involves crawling through web sites to extract URLs and then extract corresponding html doc from web pages. The second phase, extract relevant information from html doc while the third phase transform the extracted information into a structured format.

Both Web scraping and API shared similar characteristics of information extraction. However, they differed in the type and amount of information they extract. API scraping has restricted data retrieval capabilities. Web scraping can be done using programming languages such as PHP, Node JS, and Python, R etc. or readily available tools such as scrapy, import.io etc. can be utilized. Web scraping is a very efficient and important technology used in almost all field of human endeavors. Web scraping is also grappled with difficulties such as anti-scraping techniques such as CAPTCHA that prevent the scraping of certain websites. Scraper's developers occasionally overcome this obstacle but often slows down the scraping process. Furthermore, there is no clear legal consensus on whether web scraping is permissible or not. Thus, the legal landscape surrounding web scraping is still evolving.

4. Conclusion

Data scraping or web scraping as it is popularly known is a technology that is getting more and more attention due to its importance in virtually all field of human endeavor. There are many definitions of this concept, but most of the definitions revolve around the extraction and transformation of the huge amount of unstructured information found in websites into structured format to facilitate decision making. Albeit there is a slight difference between data scraping and web scraping. The former is more versatile, it supports the extraction of information from both online and offline sources, while the latter is specific to the extraction of information from website.

Web scraping, web crawling and data mining are used interchangeably. Apparently, Web Scraping and Web Crawling are similar as both refer to the collection of information from websites but differ in the type and amount of information they extract. While Data mining is an activity that takes place after Web Scraping, thus its focuses on data analysis. The phases or stages of Web Scraping is further simplified in figure 1 to enhance comprehension. This is because most studies in Data Scraping only focused on textual explanation of these phases. Moreover, the differences between Web Scraping and API Scraping were extracted from several authors and presented in tabular form to give a brief overview of the two concepts. For instance, Web scraping has more advantages when compared with API scraping since the former has little or no restriction when scraping websites. Additionally, there exist a handful of web scraping tools that can be utilize for data extraction. Most of these tools are free but has limited functionality. Furthermore, this relatively new concept is assisting different field of human endeavor extract data for analysis and decision-making process. Web Scraping is also faced with certain challenges such as the deployment of ant-scraping technology to prevent the scraping of certain websites. However, some developers always find a way of by passing such restrictions. Finally, it is legally allowed to scrape websites if the terms and conditions of such sites are respected.

References

[1] A. Oussous, F. Z. Benjelloun, A. Ait Lahcen, and S. Belfkih, "Big Data technologies: A survey," *Journal of King Saud University - Computer and Information Sciences*, vol. 30, no. 4, pp. 431–448, 2018, doi: 10.1016/j.jksuci.2017.06.001.

[2] B. Batrinca and P. C. Treleaven, "Social media analytics: a survey of techniques, tools and platforms," *AI & Soc*, vol. 30, no. 1, pp. 89–116, 2015.

[3] J. R. Gallagher and A. Beveridge, "Project-Oriented Web Scraping in Technical Communication Research," *Journal of Business and Technical Communication*, vol. 36, no. 2, pp. 231–250, 2022.

[4] R. Sharma, "Web Data Scraping," *Kalinga Institute of Industrial Technology*, no. March, p. 4, 2020.

[5] R. Vording, "Harvesting unstructured data in heterogenous business environments ; exploring modern web scraping technologies," Netherlands.

[6] B. Zhao, "Web Scraping," *Encyclopedia of Big Data*, 2017, doi: 10.1007/978-3-319-32001-4.

[7] J. Hillen, "Web scraping for food price research," *British Food Journal*, vol. 121, no. 12, pp. 3350–3361, 2019.

[8] T. Karthikeyan, K. Sekaran, D. Ranjith, V. Vinoth kumar, and J. M. Balajee, "Personalized content extraction and text classification using effective web scraping techniques," *International Journal of Web Portals*, vol. 11, no. 2, pp. 41–52, 2019, doi: 10.4018/IJWP.2019070103.

[9] C. Yaiprasert and G. Yusakul, "Artificial intelligence for target symptoms of Thai herbal medicine by web scraping," *International Journal of Data and Network Science*, vol. 6, no. 3, pp. 1013–1022, 2022.

[10] K. Henrys, "Importance of Web Scraping in E-Commerce and E-Marketing," *IT Official Journal*, no. December 2020, pp. 1–10, 2021, doi: 10.2139/ssrn.3769593.

[11] E. L. Nylen and P. Wallisch, "Web Scraping," *Neural Data Science*, pp. 277–288, 2017, doi: 10.1016/b978-0-12-804043-0.00010-6.

[12] M. A. Khder, "Web scraping or web crawling: State of art, techniques, approaches and application," *International Journal of Advances in Soft Computing and its Applications*, vol. 13, no. 3, pp. 144–168, 2021, doi: 10.15849/ijasca.211128.11.

[13] A. Bradley and R. J. E. James, "Web Scraping Using R," *Advances in Methods and Practices in Psychological Science*, vol. 2, no. 3, pp. 264–270, 2019.

[14] S. Han and C. K. Anderson, "Web Scraping for Hospitality Research: Overview, Opportunities, and Implications," *Cornell Hospitality Quarterly*, vol. 62, no. 1, pp. 89–104, 2021, doi: 10.1177/1938965520973587.

[15] E. Suganya and S. Vijayarani, *Sentiment Analysis for Scraping of Product Reviews from Multiple Web Pages Using Machine Learning Algorithms*, vol. 941. Springer International Publishing, 2020.

[16] T. G. D. Souza, F. D. R. Fonseca, V. D. O. Fernandes, and J. C. Pedrassoli, "Exploratory spatial analysis of housing prices obtained from web scraping technique," *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, vol. 43, no. B4-2021, pp. 135–140, 2021, doi: 10.5194/isprs-archives-XLIII-B4-2021-135-2021.

[17] K. Parikh, D. Singh, D. Yadav, and M. Rathod, "Detection of Web Scraping using Machine Learning," *Open access international journal of Science and Engineering*, vol. 3, no. 1, pp. 114–118, 2018.

[18] E. Persson, "Evaluating tools and techniques for web scraping," KTH ROYAL INSTITUTE OF TECHNOLOGY, 2019.

[19] L. R. Julian and F. Natalia, "The use of web scraping in computer parts and assembly price comparison," *CONMEDIA 2015 - International Conference on New Media 2015*, 2016, doi: 10.1109/CONMEDIA.2015.7449152.

[20] Manjushree B.S and S. G.S, "Survey on Web scraping technology," *Waffen-Und Kostumkunde Journal*, vol. 16, no. 06, pp. 1–8, 2020, doi: 10.37896/whjj16.06/001.

[21] P. Thota and E. Ramez, "Web Scraping of COVID-19 News Stories to Create Datasets for Sentiment and Emotion Analysis," *ACM International Conference Proceeding Series*, pp. 306–314, 2021, doi: 10.1145/3453892.3461333.

[22] S. Pai, "Web Scraping vs API: What's the best way to extract data [Blog post]." 2021.

[23] T. Paul, "10 Web Scraping Challenges You Need to Know [Blog post]." 2021.

[24] S. Chaudhari, R. Aparna, V. G. Tekkur, G. L. Pavan, and S. R. Karki, "Ingredient/Recipe Algorithm using Web Mining and Web Scraping for Smart Chef," *Proceedings of CONECCT 2020 - 6th IEEE International Conference on Electronics, Computing and Communication Technologies*, no. 3, pp. 22–25, 2020, doi: 10.1109/CONECCT50063.2020.9198450.

[25] N. R. Haddaway, "The Use of Web-Scraping Software in Searching for Grey Literature," *Grey Journal*, vol. 11, no. February, pp. 186–190, 2015.

[26] L. Richard, B. Robbie, C. Katelyn, and C. Andrew, "A Primer on Theory-Driven Web Scraping: Automatic Extraction of Big Data From the Internet for Use in Psychological Research," *Psychol Methods*, 2016.

[27] A. Maududie, W. E. Y. Retnani, and M. A. Rohim, "An Approach of Web Scraping on News Website based on Regular Expression," in *Proceedings - 2nd East Indonesia Conference on Computer and Information Technology: Internet of Things for Industry, EIConCIT 2018*, Jember, Indonesia: IEEE, 2018, pp. 203–207. doi: 10.1109/EIConCIT.2018.8878550.

[28] C. Muehlethaler and R. Albert, "Collecting data on textiles from the internet using web crawling and web scraping tools," *Forensic Science International*, vol. 322, p. 110753, 2021, doi: 10.1016/j.forsciint.2021.110753.

[29] V. Krotov and L. Silva, "Legality and ethics of web scraping," *Americas Conference on Information Systems 2018: Digital Disruption, AMCIS 2018*, no. September, p. 35, 2018.

[30] A. Luscombe, K. Dick, and K. Walby, "Algorithmic thinking in the public interest: navigating technical, legal, and ethical hurdles to web scraping in the social

sciences," *Quality and Quantity*, no. 0123456789, 2021, doi: 10.1007/s11135-021-01164-0.
[31] S. Scheps, *Business Intelligence for Dummies*. Indianapolis, Indiana: Wiley Publishing, Inc., 2008.