A Tool for Study on Impact of Big Data Technologies on firm performance

Chaimaa Lotfi[1], Swetha Srinivasan[1], *Myriam Ertz[1][0000-0001-9959-2779] and Imen Latrous [1]

[1] LaboNFC, University of Quebec at Chicoutimi, 555 Boulevard de l'Université, Saguenay (QC), Canada

[1] chaimaa.lotfi@etu.toulouse-inp.fr
[2] swethas162001@gmail.com
[3*] Myriam_Ertz@uqac.ca
[4] Imen_Latrous@uqac.ca

* Corresponding Author

Abstract. Organizations can use big data analytics to evaluate large data volumes and collect new information. It aids in answering basic inquiries concerning business operations and performance. It also aids in the discovery of unknown patterns in massive datasets or combinations of datasets. Overall, companies use big data in their systems to enhance operations, provide better customer service, generate targeted marketing campaigns, and take other activities that can raise revenue and profitability in the long run. Therefore, it's becoming increasingly important to apply and analyze big data approaches for business growth in today's data-driven world. More precisely, given the abundance of data available on the Internet, whether via social media, websites, online portals, or platforms, to mention a few, businesses must understand how to mine that data for meaningful insights. In this context, web scraping is an essential strategy. As a result, this work aims to explain the application of the developed tool to the specific case of retrieving big data information about the particular companies in our sample. The paper starts with a short literature review about web scraping then discusses the tools and methods utilized, describing how the developed technology was applied to the specific scenario of retrieving information about big data usage in the enterprises present in our sample.

Keywords: Big Data, Web Scraping, Text Mining, Data, Internet, Social Media, Information, Business.

1 Purpose and objectives

Companies use big data in their systems to enhance operations, provide better customer service, generate targeted marketing campaigns, and take other activities that can raise revenue and profitability in the long run. As a result, businesses who properly use it have a potential competitive advantage over those that don't since they can make more informed and faster business decisions [1].

Big data provides firms with important customer insights that they can utilize to improve their marketing, advertising, and promotions in order to boost customer engagement and conversion rates. Consumer and corporate purchasers' developing preferences can be assessed using both historical and real-time data, allowing organizations to become more responsive to their wants and needs [1].

Every day roughly 2.5 quintillion bytes of data are generated globally [2]. If this data is not turned into insights, the whole process of generating data becomes futile. Big data requires new technical architectures, analytics, and tools to enable insights that unlock new sources of corporate value due to its magnitude, spread, diversity, and/or timeliness. Big data is defined by three primary characteristics: volume, variety, and velocity, or the three V's. The extent and scope of the data are determined by its volume. The rate at which data changes or is created is called velocity. Finally, variety refers to the various formats and types of data, as well as the various us es and methods of data analysis [3].

Firms like Google, eBay, LinkedIn, and Facebook were built around big data from the beginning [4]. Firms in the retail sector like Walmart established Walmart Labs and Fast Big Data team to help determine the price, improve customer experience and solve problems rapidly. Entertainment houses like Netflix employ BDA techniques to improve user experience by introducing accurate recommendation algorithms.

In this paper, we have developed a Web scraping tool to study how the firms use BDA techniques to improve their performance based on information available in academic literature. There are various BDA techniques. For a more extensive and exhaustive analysis, we chose INFORMS (Institute for Operations Research and Management Science) classification consisting of three groupings, namely descriptive, predictive and prescriptive analysis [5,6].

- Descriptive Analytics deals with report generation activities in order to answer the question of “what happened?” or “what is happening?”. It includes techniques, namely Association Analysis, Sequence Analysis, Cluster Analysis, Similarity Matching, Link Analysis.
- Predictive Analytics refers to the process of estimating or forecasting future events to answer the question “what will happen.” Decision Trees, Neural Networks, Partial Least Squares Regression, least-angle regression (LARS) come under this group.
- Prescriptive Analytics helps make decisions based on analysis made by descriptive and predictive methods to answer the question of “what should be done?”. It consists of Stochastic Optimization, Optimization, Multi-Criteria Decision Making Techniques, Decision Modeling, Network Science, Simulation Techniques, Deep Learning, Artificial Intelligence.

The firms studied in this research are part of the S&P 500 (Standard & Poor's 500) in the USA and the S&P/TSX 60 (Standard & Poor's/Toronto Stock Exchange 60) in Canada.

The paper follows the regular Web scraper design procedure as its conceptual background. The conceptual framework can be summarized as follows in Figure 1.

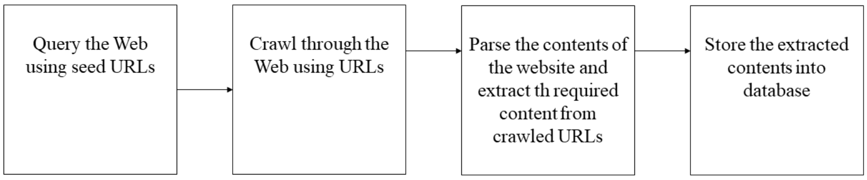
Query the Web using seed URLs
Crawl through the Web using URLs
Parse the contents of the website and extract th required content from crawled URLs
Store the extracted contents into database

Fig. 1. The basic design of a web scraper (Source: The authors)

Figure 1 depicts the overarching guiding schema that we will follow subsequently in our study. Overall, from left to right, we start by querying the Web using seed URLs. Afterward, we crawl through the Web using seed URLs. Thirdly, we parse the contents of the website and extract the required content from crawled URLs. Finally, the extracted content is stored in a comprehensive database.

Section 2 outlines the methodology, whereas Section 3 outlines the paper's main result and contribution, which is the proposed methodology to develop the Web scraping tool. The section also summarizes the key findings of the study and the potentialities to improve the tool even further. Finally, Section 4 wraps up the paper and provides future research avenues.

2 Literature review

The topic of Big Data analytics has sparked a lot of interest from the researcher and practitioner communities. The literature provides ample evidence of the many ways in which Big Data analytics contributes to improving operational efficiency while improving customer service, to name but a few of their benefits.

In terms of improving operations, Big Data Analytics (BDA) plays a key role in Inventory Management (IM), Supply Chain Management (SCM), and Logistics Management (LM) because it analyzes customer behavior and uses the knowledge derived from those analyses to optimize business operations [7]. The recourse to artificial intelligence (AI) and Big Data also proved to assist companies in responding to the COVID-19 [8]. To provide a summary from the ever-growing literature on Big Data analytics in operations, Talwar et al. [9] develop the Dimensions-Avenues-Benefits (DAB) model as a conceptual framework summarizing the key themes, existing research trends, and future research avenues for BDA adoption in operations and supply chain management. BDA might further benefit organizations in their efforts to become more sustainable. Chandra and Verma [10] suggest that BDA assists in sensing customer desire for companies to develop green operations and supply chains. This type of customer knowledge will be conducive to altering business models as well [10].

Regarding customer service improvement, researchers have sought many ways to improve human-human relationships in business settings and human-computer interactions. For example, Kottursamy [11] uses different deep learning analysis algorithms for emotion recognition and sentiment analysis purposes in an attempt to improve interactions with customers. To delve more specifically into predicting children's behavior according to their emotions, Kumar and Senthil [12] propose using the deep learning classifiers methods. They suggest a fusion of the Naïve Bayes method and decision tree algorithm for greater accuracy. BDA might also offer interesting solutions for customer service in e-commerce. In fact, businesses face contradictory signals in that they need to provide excellent customer service, but they also face high turnover rates in customer service departments, high labor costs, and difficult recruitment. One solution resides in implementing automatic customer service. Kong and He [13] propose a customer service system based on big data machine learning to better serve consumers' needs more efficiently and conveniently. Furthermore, with the continuous increase in user-generated content (UGC), consumers express their concerns explicitly, and firms may use text mining capabilities to enhance product attributes and service characteristics accordingly [14]. To Kitsios et al. [14], these applications thus support marketing activities and improve business value.

Recommendations are also a very common way in which big data analytics can improve consumer service. They assist users in sifting through a vast quantity of seemingly undifferentiated offers to select the best-fitting or most appropriate one according to consumer preferences. Recent research has emphasized that location and orientation are highly influential in determining consumer preferences more efficiently. Accordingly, in addition to content-based recommender algorithms and collaborative filtering, Joe and Raj [15] developed a "location-based orientation context-dependent recommender system" (p. 14) that is fed with user data obtained from IoT devices (e.g., Google Home, iPhones, smartwatches, smartphones) to account for the user context. In the healthcare sector, recommendation systems are also increasing in accuracy and security through the use of advanced deep learning classifiers. For example, Manoharan [16] proposes a "K-clique embedded deep learning classifier recommendation system" (p. 121) for improving automatic diet suggestions to patients according to their health conditions (e.g., sugar level, age, cholesterol level, blood pressure).

Meanwhile, the gathered data also suffers from the likelihood of being exposed to the public or exploited. Some researchers have therefore sought ways to protect data. For example, Haoxiang and Smys [17] propose a perturbation algorithm using Big Data employing "optimal geometric transformation" (p. 19) for higher scalability, accuracy, resistance, and execution speed compared to other privacy preservation algorithms.

3 Methodology

Data has an essential role in business, marketing, engineering, social sciences, and other fields of study since it may be utilized as a starting point for any operations involving the utilization of information and knowledge. The initial stage of research is data collecting, followed by systematic assessment of information about essential elements, which allows one to answer questions, design research questions, test hypotheses, and evaluate outcomes. Several data collection methods are used depending on the subject or topic of study, the nature of the information sought, and the user's goals or objectives. Depending on the purposes and situations, the application methodology can also change without impacting data integrity, correctness, or reliability [18]. Several data sources on the Internet can be used in the research process. Web scraping, web extraction, web harvesting, and web crawling are all terms used to describe the process of obtaining data from websites. This study will look into how to develop a web scraping tool for extracting valuable data from internet sources and the most recent advancements in web scraping approaches and techniques. This tool will help study how the firms use BDA techniques to improve their performance based on information available in academic literature. The research also assisted us in comparing the various tools available and choosing the most suitable one for the study.

2.1 Development of Web Scraping tools

The web data extraction tool can be tailor-made for each specific application. The following section discusses how the web data extraction tool has been built using different techniques.

2.1.1 Web Scraping using BeautifulSoup

Beautiful Soup is a Python library that allows parsing HTML and XML files. In addition, it builds a parse tree for parsed pages, which may be used to extract data from HTML and is helpful for web scraping.

2.1.2 Web Scraping using Java libraries

Crawler4j is an open-source Java crawler with a simple user interface for indexing web pages. This software makes it simple to build up a multi-threaded crawler in a short amount of time.

2.1.3 Web Scraping using Selenium

Selenium is an open-source web-based automation tool that is quite efficient at scraping websites. Selenium's web driver has several features that allow users to move across web pages and retrieve different page parts depending on their needs.

As a result, many data from several websites related to the user's query can be retrieved and organized.

2.1.4 Web Scraping using Apache Nutch

Apache Nutch is an open-source large-scale distributed web-crawler developed in Java language that can be extended very easily.

2.1.5 Web Scraping using Scrapy

Scrapy (SKRAY-pee) is a Python-based web crawling framework that is free and open-source. It was created with web scraping in mind, but it may also collect data via APIs or as a general-purpose web crawler.

2.1.6 Web Scraping using R

The RCrawler can crawl, parse, store, and extract material from online sites, as well as generate data that may be used directly in web content mining applications. Multi-threaded crawling, content extraction, and duplicate content detection are the core characteristics of RCrawler.

2.1.7 Comparison of web scraping tools

Table 1 summarizes and compares the different tools that can be used for web scraping purposes. Table 2 focuses more specifically on the Python Web scraping libraries and frameworks.

Table 1. Comparison between Open-Source Web scraping techniques and frameworks

Parameters	Type[1]	API/standalone	Language	Extraction facilities[2]
Jsoup	CP	API	Java	H, C
HttpClient	C	API	Java	
Scrapy	F	Both	Python	R, X, C
BeautifulSoup	P	No	Python	H
Apache Nutch	F	Both	Java	R, X, H, C
Selenium	P	API	Java, Python	R, X, C

Notes: [1] Type: C = HTTP Client

[2] Extraction facilities:

R = Regular expressions   P = Parsing   H = HTML parsed tree
F = Framework   X = XPath   C = CSS selectors

Table 2. Comparison between Python Web scraping libraries and frameworks

Factors	BeautifulSoup	Scrapy	Selenium
Extensibility	Suitable for low-level complex projects	Best choice for large or complex projects	Best for projects dealing with Core JavaScript
Performance	Pretty slow compared to other libraries while performing a specific task	Rapid processing due to the use of asynchronous system calls	Can handle up to some level but not as much as Scrapy
Ecosystem	Has a lot of dependencies on the ecosystem	Has a flexible ecosystem making it easy to integrate with proxies and VPNs	Has a good ecosystem for the development

Based on the analysis above, it is identified that Scrapy is a better framework than others for extensively performing web data extraction. Following are research studies that use the Scrapy framework for their research purpose.

Muehlethaler and Albert [18] used Scrapy and Kibana/Elastic search interface to collect data from online clothes retailers by scraping its website. They successfully extracted 68 text-based fields describing a total of 24,701 clothes to help provide precise estimations of fibers types and color frequencies within a timeframe of 24 hours.

Seliverstov et al. [19] developed a web scraper using the Scrapy framework to collect published reviews, which was further processed using Natural Language Processing techniques to classify the reviews. The quality of roads in the Northwestern Federal district was evaluated based on the classification.

Shen et al. [20] employ the Scrapy and Common Crawl framework to collect text data from the Internet for the proposed model to classify food images and estimate food attributes accurately.

Maroua et al. [21] developed a tool named WebT-IDC, Web Tool for Intelligent Data Creation which can construct noiseless corpora of feedback on different topics by extracting relevant data from web forums and blogs using the Scrapy framework.

Budiarto et al. [22] leverage the Scrapy framework to extract news articles from specific portals and extract latent information using topic modeling and spherical clustering.

4 Results and findings

4.1 Data Collection:

With the pre-constituted corpus of publications collected in 2019 as the basis, we collected data from academic literature with the help of a web scraping tool and built an updated corpus.

To update the corpus to include papers from 2020 and 2021, we collected papers that cite the papers already present in the corpus using a web scraping tool. The process of updating the corpus is presented as two modules:

- Extracting cited papers
- Downloading the cited papers

4.1.1 Extracting cited papers

The pre-constituted corpus had publications related to the impact of big data analytics on firm performance. We used the tool to identify publications that cite these research papers to update the corpus with relevant and recent research studies on big data analytics capabilities [23,24]. We leveraged Google Scholar to identify the citing publications. We built a tool that can search for the title of a research paper in the publication and extract titles that cite that paper. Figure 2 explains the structure of the tool developed. As shown from top to bottom, we first get a list of titles from the CSV file. Then, we open Google Scholar with a proxy for each title and enter the title in the search bar. We then select the following: "cited by" option and "since 2020" timeframe. For each page in the results page, we scrap the title and the link using XPath and CSS Selectors of Scrapy. We then save the scraped data in the form of a dictionary and append it for each page iteration. Finally, we convert the dictionary into a data frame and save the data frame as a CSV file. This gives the CSV file of the scraped data as the output.

Selenium is an open-source web-based automation tool that is quite efficient for web scraping and was employed to scrape data from Google Scholar. The web driver in Selenium provides numerous features that enable users to navigate the desired web pages and fetch various page contents depending on necessities [25]. Thus, many data from multiple web pages concerning the user's query can be extracted from numerous web pages and grouped. However, during scraping, issues concerning a CAPTCHA and integrating Selenium with proxies and VPNs were faced [26].

Octoparse, a cloud-based web data extraction solution that helps users extract relevant information from various websites, was also used, but the tool did not provide the required results [27]. Finally, ScraperAPI, an API that handles proxies, browsers, and CAPTCHAs to avoid triggering CAPTCHA, was used along with Scrapy, which supports integration with proxies and VPN [28].

The tool was built based on Scrapy architecture [29]. The tool was used to navigate the page to provide accurate results. The tool initially gets the titles from a CSV file. For each title in the file, the tool opens Google Scholar, enters the title in the search box, selects the "Cited by" option to identify papers that cite these publications, and chooses the year "2020" to identify more recent publications. This will lead Google Scholar to give us recent and updated publications that cite the research papers in the pre-constituted corpus. From the results obtained, the tool will further identify the title of each document that cites the publication and extract the title and link to each of the papers with which the document can be downloaded. We employ Scrapy's CSS and XPath selectors to identify the title and link of each citing publication and extract those results in the form of a CSV file. The following figure depicts the process flow of the tool (see Figure 2).

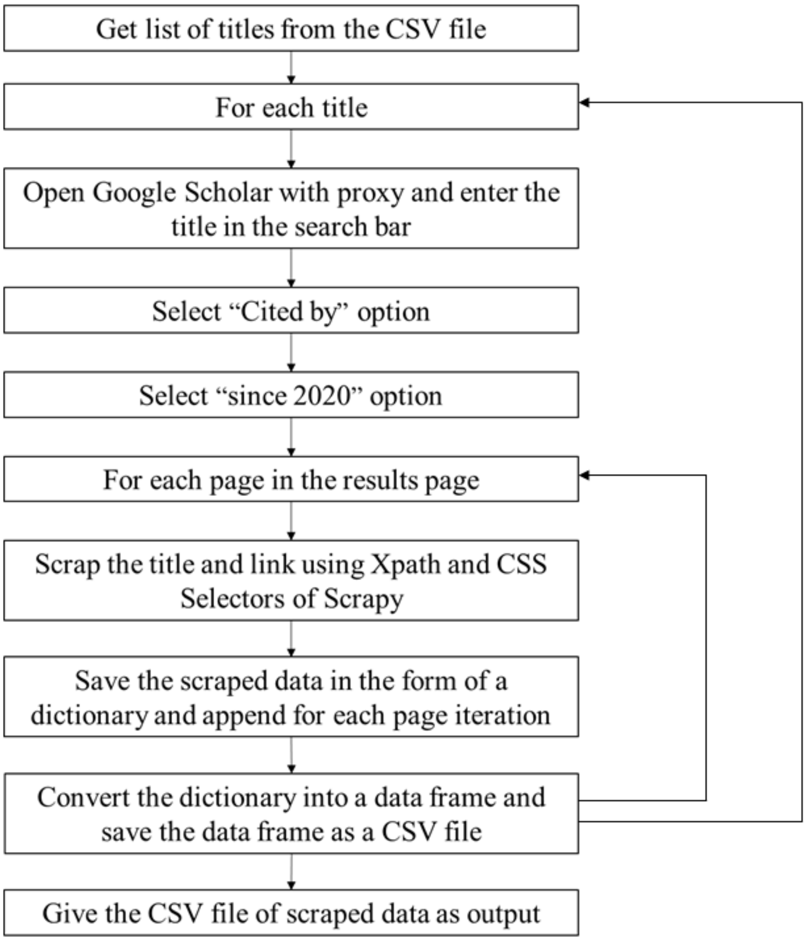
Get list of titles from the CSV file
For each title
Open Google Scholar with proxy and enter the title in the search bar
Select "Cited by" option
Select "since 2020" option
For each page in the results page
Scrap the title and link using Xpath and CSS Selectors of Scrapy
Save the scraped data in the form of a dictionary and append for each page iteration
Convert the dictionary into a data frame and save the data frame as a CSV file
Give the CSV file of scraped data as output

Fig.2. Process flow of the web scraping tool (Source: The authors)

3.1.2 Downloading the cited papers

In order to download the cited papers, the following procedure has been implemented:

1. Open the CSV file of cited documents from the previous module
2. Convert the contents of the CSV file into a data frame
3. For each link in the data frame
    - Identify the links with PDF extension
    - Enter the link
    - Click on save

To download the cited papers, with the help of the extensions of the link, they were classified as PDF files and other links. If the links are directly linked to a PDF file,

two libraries - Requests and Mimetypes - are leveraged to extract the file. The "requests" library is an HTTP library for Python, with its method GET to retrieve the PDF from the link provided. The second library is a module that converts between a URL and the mime-type associated with the filename extension and helps identify if a link has a PDF extension or not [24].
We download the paper if the link's extension is PDF during the next step, using a simple code with requests library. Otherwise, we keep the link in a CSV file to download it manually.
If the link extension is not PDF, we manually download each paper using sci-hub if the link is not accessible.

4.2 Data Processing

Once the corpus is updated, a text-mining tool is built to help us extract meaningful insights from the data collected. This process is further divided into two modules:

- Converting PDF to text
- Searching for companies' names and technique names

4.2.1 Converting PDF to text

The papers are stored in Google Drive to deal with storage constraints and access constraints. The documents stored in the drive are accessed individually in PDF format. We parse those files individually and store the contents in a text file format in Google Drive. We use the Apache Tika library to parse the file content and store them in a text file [26]. Apache Tika is a library used for document type detection and content extraction from various file formats. Using this, one can develop a universal type detector and content extractor to extract both structured text and metadata from different types of documents such as spreadsheets, text documents, images, PDFs, and even multimedia input formats to a certain extent.

4.2.2 Searching for companies' names and technique names

After converting all the PDFs into text files, we use a simple function that searches for the company's name in the text file. If the company's name is found, we start searching for the technique name, and then if both are located, we print the matched lines of the document. Following this, we manually read these lines and tried to figure out if that company was using that particular technique.

We have also tried implementing spaCy library, an open-source NLP library. Matcher, a rule-matching engine featured by spaCy, would have helped us find what we are looking for in the text document. It works better than regular expressions because the rules can also refer to annotations. However, the tool needs to be better refined to acquire accurate results.

4.2.3 Summary of key findings and challenges

This section summarizes the key findings of the study and the challenges encountered. More specifically:

- We were able to find a lot of company names and technique names, but both were in different contexts.

- We could find a lot of familiar and popular company names in the academic literature.

- Generic technique names like "artificial intelligence" or "optimization" or "descriptive," "predictive," and "prescriptive analytics" or just "analytics" or "big data" appeared a lot more than specific technique names

Overall, the proposed Web scraping tool is reasonably precise and helpful. Still, it will potentially need further refinements to scrape for more specific big data technique names above and beyond the generic names. Besides, the tool can connect information that emanated from disjoint contexts, and this capacity could even be strengthened further by increasing the power of the tool to large aggregate amounts of data originating from more diverse contexts and in different languages.

The following figures depict the most popular techniques in academic literature. Figure 3 indicates that clustering/cluster analysis was the technique used by many companies. Under predictive analytics (see Figure 4), neural networks, and to a lesser extent, decision trees emerge as popular techniques. Finally, generic techniques like “deep learning” and “optimization” seem to prevail among companies under prescriptive analytics. (see Figure 5).

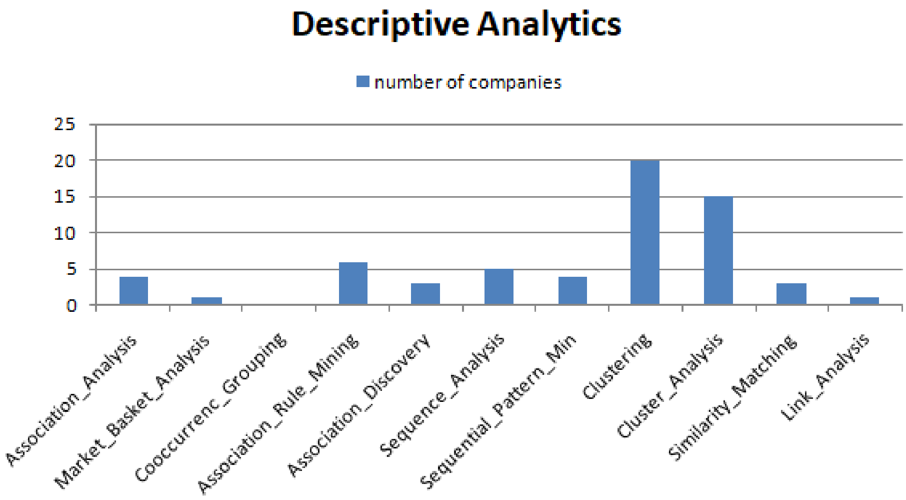

Fig.3. No. of companies using Descriptive Analytics Techniques

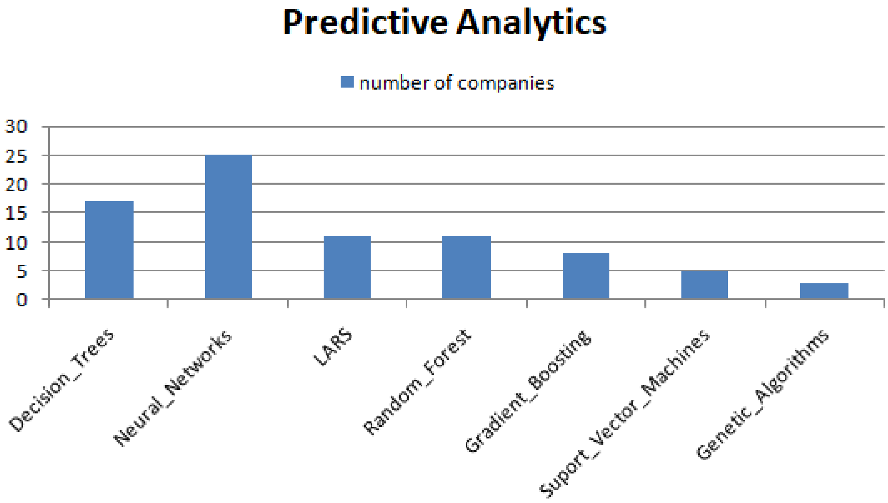

Fig.4. No. of companies using Predictive Analytics Techniques

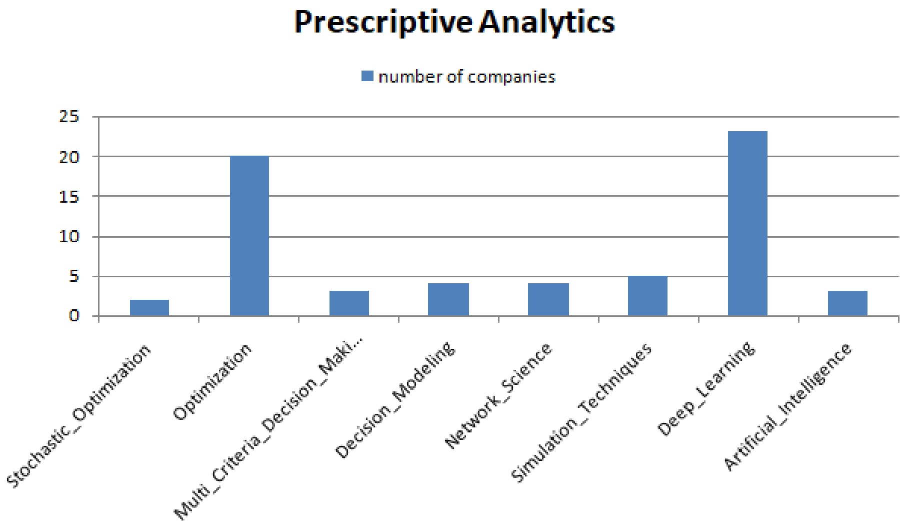

Fig.5. No. of companies using Prescriptive Analytics Techniques

This study makes a strong research case in that the proposed Web scraping tool is reasonably precise and helpful. Still, it will potentially need further refinements to scrape for more specific big data technique names above and beyond the generic names explored in the framework of this study. Besides, the tool can connect information that emanated from disjoint contexts, and this capacity could even be strengthened further by increasing the power of the tool to large aggregate amounts of data originating from more diverse contexts and in different languages.

5 Conclusion and Future Scope

In this study, we analyzed the recent literature on web scraping applications in various areas, web scraping methodologies, and tools that use web scraping techniques. We discovered that Scrapy delivers better results when evaluating the performance and functionality of different tools and frameworks since it is fast, versatile, and powerful. Scrapy processes requests asynchronously, so the results may be scraped quickly. Scrapy's design is built around a web crawler that makes data extraction simple. Based on the results provided by Scrapy, we further developed a tool to identify and extract information relevant to the use of big data techniques by firms. This study demonstrates how the developed tool can be applied to the specific instance of retrieving big data information about the organizations in our sample.

With this Web scraping tool, researchers and data scientists might extract valuable information from online databases or repositories for further applications such as studying the relationship between big data usage and firm performance. More specifically, future research might use the proposed tool in order to determine the extent to which big data analysis can improve a company's financial performance. In fact, according to Erevelles et al. [30], Big Data enables enterprises to collect and keep more exact and accurate information on a variety of areas of their businesses, including their impact on performance. As a result, it provides a more accurate picture of the customer, leading to better consumer and marketing information [31]. Enhanced analytical capabilities will lead to better marketing tactics and, in turn, improved business performance.

Furthermore, future research might use the proposed tool in order to determine what conditions lead to big data making a more significant contribution to the company. The current tool might be used in a three-stage project to accomplish this. The project's initial stage is applying the proposed Web scraping tool to examine a large corpus of academic and professional literature on the usage of Big Data Analytics by large corporations. The second stage would involve using the Web scraping tool in order to extract useful information on the selected companies' use of Big Data Analytics from the corpus of texts. In addition, to supplement BDA insights, financial performance metrics need to be retrieved from private databases (e.g., Mergent, Orbis). Finally, an econometric model might be constructed during the third stage to examine the influence of big data analytics on corporate performance.

Acknowledgments

The authors are grateful to the MITACS GLOBALINK program for providing valuable financial and non-financial support in the conduct of this project.

References

1. Botelho, B. “Editorial Director, News - TechTarget – SearchEnterpriseAI” (2022) https://www.techtarget.com/contributor/Bridget-Botelho
2. SeedScientific. “How Much Data Is Created Every Day? [27 Staggering Stats]” October 28 (2021) https://seedscientific.com/how-much-data-is-created-every-day/
3. Elgendy, N., Elragal, A.: “Big Data Analytics: A Literature Review Paper. Industrial conference on data mining.” Springer, Cham (2014): 214-227. https://doi.org/10.1007/978-3-319-08976-8_16
4. Davenport, T.H., Dyché., J.: "Big data in big companies." International Institute for Analytics, 3 (2013): 1-31.
5. Delen, Dursun, Sudha Ram. "Research challenges and opportunities in business analytics." Journal of Business Analytics 1.1 (2018): 2-12. https://doi.rog/10.1080/2573234X.2018.1507324
6. Ertz, M., Sun, S., Latrous, I.: "The Impact of Big Data on Firm Performance." International Conference on Advances in Digital Science. Springer, Cham (2021): 451-462. https://doi.org/10.1007/978-3-030-71782-7_40

7. Maheshwari, Sumit, Prerna Gautam, and Chandra K. Jaggi. "Role of Big Data Analytics in supply chain management: current trends and future perspectives." International Journal of Production Research 59.6 (2021): 1875-1900.
8. Chen, Yasheng, and Mohammad Islam Biswas. "Turning Crisis into Opportunities: How a Firm Can Enrich Its Business Operations Using Artificial Intelligence and Big Data during COVID-19." Sustainability 13.22 (2021): 12656.
9. Talwar, Shalini, et al. "Big Data in operations and supply chain management: a systematic literature review and future research agenda." International Journal of Production Research (2021): 1-26. https://doi.org/10.1080/00207543.2020.1868599
10. Chandra, S., Verma, S. (2021). Big data and sustainable consumption: a review and research agenda. Vision, https://doi.org/10.1177/09722629211022520
11. Kottursamy, Kottilingam. "A review on finding efficient approach to detect customer emotion analysis using deep learning analysis." Journal of Trends in Computer Science and Smart Technology 3.2 (2021): 95-113.
12. Kumar, T. Senthil. “Construction of Hybrid Deep Learning Model for Predicting Children Behavior based on their Emotional Reaction.” Journal of Information Technology 3, no. 01 (2021): 29-43.
13. Kong, Yuqiang, and Yaoping He. "Customer Service System Design Based on Big Data Machine Learning." Journal of Physics: Conference Series. Vol. 2066. No. 1. IOP Publishing, 2021.
14. Kitsios, Fotis, et al. "Digital marketing platforms and customer satisfaction: Identifying eWOM using big data and text mining." Applied Sciences 11.17 (2021): 8032.
15. Joe, Mr C. Vijesh, and Jennifer S. Raj. "Location-based Orientation Context Dependent Recommender System for Users." Journal of trends in Computer Science and Smart technology (TCSST) 3.01 (2021): 14-23.
16. Manoharan, Samuel. "Patient diet recommendation system using K clique and deep learning classifiers." Journal of Artificial Intelligence 2.02 (2020): 121-130.
17. Haoxiang, Wang, and S. Smys. "Big Data Analysis and Perturbation using Data Mining Algorithm." Journal of Soft Computing Paradigm (JSCP) 3.01 (2021): 19-28.
18. Muehlethaler, Cyril, and René Albert. "Collecting data on textiles from the internet using web crawling and web scraping tools." Forensic Science International 322 (2021): 110753.
19. Seliverstov, Yaroslav, et al. "Traffic safety evaluation in Northwestern Federal District using sentiment analysis of Internet users’ reviews." Transportation research procedia 50 (2020): 626-635.
20. Shen, Zhidong, et al. "Machine learning based approach on food recognition and nutrition estimation." Procedia Computer Science 174 (2020): 448-453.
21. Maroua, Boudabous, and Pappa Anna. "WebT-IDC: A Web Tool for Intelligent Dataset Creation A Use Case for Forums and Blogs." Procedia Computer Science 192 (2021): 1051-1060.
22. Budiarto, Arif, et al. "Unsupervised News Topic Modelling with Doc2Vec and Spherical Clustering." Procedia Computer Science 179 (2021): 40-46.

23. Suganya, E., and S. Vijayarani. "Firefly Optimization Algorithm Based Web Scraping for Web Citation Extraction." Wireless Personal Communications 118.2 (2021): 1481-1505.
24. Rahmatulloh, Alam, and Rohmat Gunawan. "Web Scraping with HTML DOM Method for Data Collection of Scientific Articles from Google Scholar." Indonesian Journal of Information Systems 2.2 (2020): 95-104.
25. Gunawan, Rohmat, et al. "Comparison of web scraping techniques: regular expression, HTML DOM and Xpath." International Conference on Industrial Enterprise and System Engineering (IcoIESE 2018) Comparison. Vol. 2. 2019.
26. Tiwari, G.: "How to handle CAPTCHA in Selenium," BrowserStack, June 8 (2021). https://www.browserstack.com/guide/how-to-handle-captcha-in-selenium
27. Octoparse, https://www.octoparse.com/
28. ScraperAPI, https://www.scraperapi.com/
29. Asikri, M. El, S. Krit, and H. Chaib. "Using Web Scraping In A Knowledge Environment To Build Ontologies Using Python And Scrapy." European Journal of Molecular & Clinical Medicine 7.3 (2020): 433-442.
30. Erevelles, Sunil, Nobuyuki Fukawa, and Linda Swayne. "Big Data consumer analytics and the transformation of marketing." Journal of business research 69.2 (2016): 897-904.
31. Hofacker, Charles F., Edward Carl Malthouse, and Fareena Sultan. "Big data and consumer behavior: Imminent opportunities." Journal of consumer marketing 33.2 (2016): 89-97.