iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VI **Month of publication:** June 2022

DOI: https://doi.org/10.22214/ijraset.2022.44841

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VI June 2022- Available at www.ijraset.com

Web Scraping for E-Commerce Websites

Gandhe Vineeth Kumar[1], Hema M S[2], Aishwarya R[3], K R Mamatha[4]
[1, 2, 3]*Undergraduate Student,* [4]*Assistant Professor, BMS College of Engineering, Bengaluru, India*

Abstract: ***The prices of the products in the E-commerce sites change frequently. It becomes very difficult for the users to monitor the prices and get the best deal available on the internet. The proposed model tackles this problem by creating a user-friendly model using Web scraping and machine learning concepts so that the model can be used by the users to monitor and compare the prices of products across the websites, send an email alert notification when there is a price drop and also to predict the future prices.***

Index terms: Web scraping, Web Crawling, BeautifulSoup, Random Forest Regression, Lasso Regression

I. INTRODUCTION

Web Scraping is a process which is used to extract data from websites which can further be used based on the requirements. It is a technique where large amounts of data can be extracted and stored in local machines in required format. Web scraping reduces the time and effort required to scrape data from the internet. The data collected from web scraping can be used in various applications like sentiment analysis, machine learning predictions and classifications, and aggregation of information to a single platform which makes it easier for accessing at one place. There are an increasing number of e-commerce sites where the prices of the products keep changing. Any user who wishes to buy a product from e-commerce sites will have to go through a number of websites in order to compare the product in all the websites. The user should select the website to buy the product such that the price is reasonable and ratings are good when compared with other websites. It becomes tedious for anyone who wants to check this manually every time they want to buy anything online. The main objective of this paper is to make the above mentioned process user friendly such that they should be able to get the best e-commerce site from which they can buy the product.

This paper comes with an approach using Web Scraping and Machine Learning to tackle these user problems and predict the price of the products and the best website from which the product should be bought. It also alerts the user through email if there is a price drop below a certain threshold.

II. RELATED WORK

A. Background Work

Numerous scrapers have been written in various programming languages and frameworks are being used for retrieving web data. Such as BeautifulSoup, Scrapy, Java, and Ruby. BeautifulSoup is used to extract banner ads from different websites [1]. Some studies explain the techniques of web scraping such as Hadul Hafeez, et.al. [2] work implemented the scraper software that is capable of collecting the updated information from the target products hosted in fabulous online e-commerce websites. Other studies discussed tools and techniques that could be used to run web scraping [3] [4] [5]. Most of these are free of cost and easy to use.

Extracting the data from an E-commerce website, based on the automatic generation of data records and summarizing the content of the entire website is defined in [6]. This is obtained by using web scraping and optical character recognition, followed by a number of nontrivial text mining and feature engineering steps. The web scraping techniques are mostly done by creating programs that automatically run queries to the web server, requesting data (usually in HTML and other forms of web pages), then parses the data to extract the necessary information and to analyze the weather related analysis in South Sumatera[7].

Marco Scarno, et.al [8] investigated the possibilities of structuring data from different websites through web scraping techniques and exploited what is offered by some web search engines to progressively create queries that enabled them to select the most useful information they needed. Some of the studies discussed the use of web scraping to extract the user information from Instagram to study and improve the features of the platform and user experience in the social media[9], [10]. While the other studies involve extraction of data in the news reporting analysis [11] and evaluating the future stock value assets [12] and Bitcoin fluctuating values in [13].

Web scraping can be automated by keeping the scheduler without burdening the extractor [14] [15]. Alvin Chandra, et.al [16] improved the social media platforms by using their respective API and Regex in the web scraping techniques. Web scraping techniques in [17] explained the complete detail for text analysis by extracting only the required text and using the Jaro-Winkler algorithm.

©IJRASET: All Rights are Reserved

The study[18] tells us that the setup of the interface used web scraping techniques along with the python modules to link a researcher's list of publications present on Google Scholar websites. While the study [19] [20] explains the price comparisons between the products extracted from the Web scraping techniques.

B. Dataset Description

The dataset plays a crucial role in training the machine learning algorithms. Scraped data of selected e-commerce sites is used for the proposed model. Dataset consists of different features like product name, price, ratings, website, timestamp. Each product is assigned a unique ID to identify the product and the e-commerce site it belongs to.

Four E-commerce websites were selected and 125 products of the electronic department were used for the experimentation. Dataset consists of a month of scraped data across selected four E-commerce sites. There are around 7716 rows of the products data consisting of varying prices in the four websites for every day. The extraction of the data is manual. Fig 1 shows features that are used from the scraped data and Fig 2 shows the correlation of the features used for training.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7716 entries, 0 to 7715
Data columns (total 10 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   product_name   7716 non-null   object
 1   price          7716 non-null   float64
 2   ratings        7716 non-null   float64
 3   website        7716 non-null   object
 4   date           7716 non-null   datetime64[ns]
 5   unq_id         7716 non-null   int64
 6   year           7716 non-null   int64
 7   month          7716 non-null   int64
 8   day            7716 non-null   int64
 9   Day_dayofweek  7716 non-null   int64
dtypes: datetime64[ns](1), float64(2), int64(5), object(2)
memory usage: 602.9+ KB
```

Fig 1: Dataset Information

	price	ratings	unq_id	year	month	day	Day_dayofweek
price	1.000000	0.016372	0.013040	nan	0.050440	-0.052937	0.014454
ratings	0.016372	1.000000	-0.703622	nan	-0.091846	0.057080	-0.057656
unq_id	0.013040	-0.703622	1.000000	nan	0.234986	-0.181060	0.152823
year	nan	nan	nan	nan	nan	nan	nan
month	0.050440	-0.091846	0.234986	nan	1.000000	-0.848480	0.388343
day	-0.052937	0.057080	-0.181060	nan	-0.848480	1.000000	-0.220563
Day_dayofweek	0.014454	-0.057656	0.152823	nan	0.388343	-0.220563	1.000000

Fig 2: Correlation matrix for the Dataset

C. Proposed system for web scraping

First, using different web scraping libraries like BeautifulSoup, Selenium and web driver, the information of the products from the selected websites is scraped. The scraped raw data is converted into readable format and stored into the database. Fig 3 shows the sample of raw scraped data.



Fig 3: Extracted raw data sample

Next, the data from the database is preprocessed into a pandas dataframe. The preprocessed data is used to train the machine learning algorithms. Variety of regression algorithms were used and evaluated on the basis of different performance metrics. The model which shows the best results is selected.

Lastly, the selected algorithm is used for predicting the future prices of the product and displayed on the dashboard. Fig 4 explains the workflow of the proposed system.

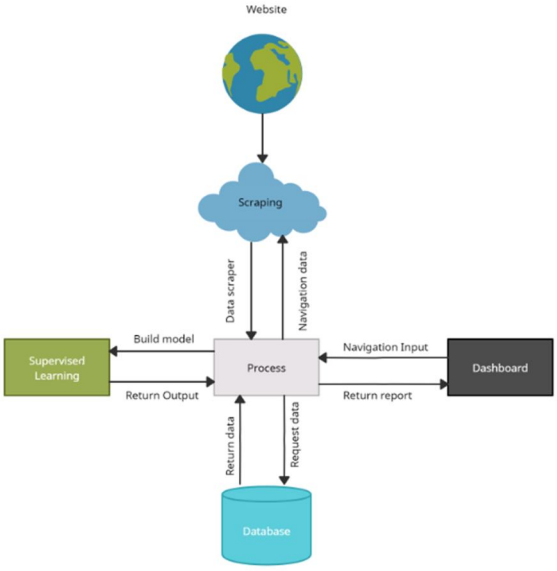


Fig 4: Block diagram of proposed model

III. MPLEMENTATION AND EXPERIMENTAL RESULTS

The project was implemented in two phases: Scraping the websites and using the scraped data into Machine learning models for product price prediction.

1) Phase 1

In phase 1, the required product items are listed out and used in the web scraping process. The web scraping is done on the selected websites by importing the necessary libraries. The scraped data will be in the form of raw data which needs to be processed thoroughly in order to get readable data as shown in Fig 5. Once the data is ready, it will be stored and used by machine learning algorithms in the further phase.

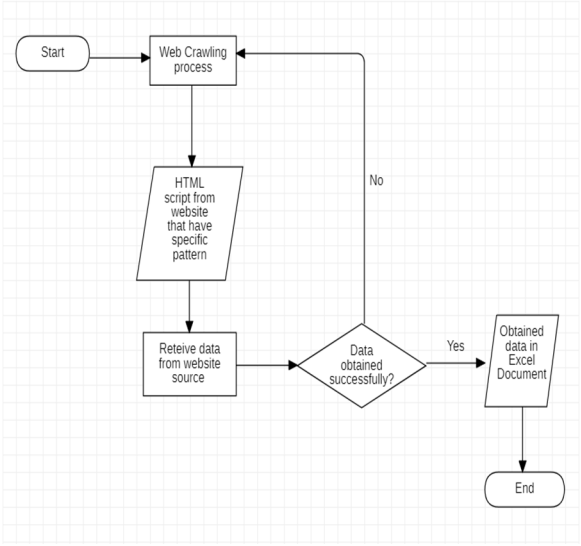


Fig 5: Phase 1 Data Flow Diagram

2) Phase 2

In the second phase, the data generated from web scraping is used. Preprocessing and sampling is performed on the dataset. Once the data is in the required format, it is splitted into a train set and test set. Different features are tried such as day of week, time stamp, ratings etc. Variety of supervised regression algorithms, like Polynomial regression, Lasso regression, SVM and Random forest regression were used while training the model. Detailed stepwise process is shown in the Fig 6

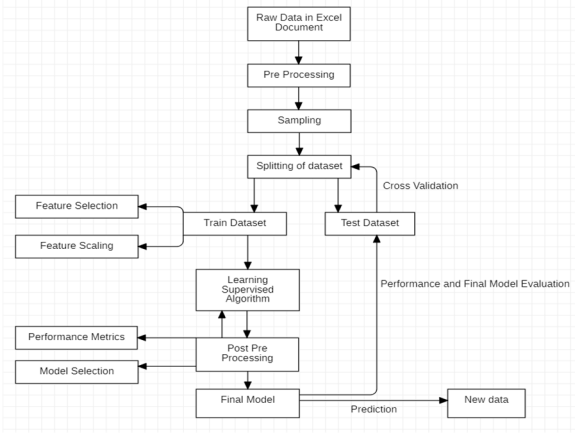


Fig 6: Phase 2 Data Flow Diagram

The evaluation of the model is performed on the test set based on the R-squared, RMSE(Root mean square error) and MAE(Mean absolute error). Table 1 shows the metrics comparison between the algorithms. It is observed that Random forest regression showed good results in terms of R-squared and SVM model showed negative R-squared value. Though the RMSE value is high for Random forest, it is observed that, with increase in days and data, the RMSE value decreased. Since Random forest outperformed remaining algorithms, it was selected for later processes such as carrying out predictions.

Table 1 : Performance metrics comparisons

	Polynomial Regression Model		**Random Forest Model**		**Lasso Regression**		**SVM**	
Metrics	Train data	Test data	Train data	Test data	Train data	Test data	Train data	Test data
RSquare	**0.008**	**0.002**	**0.99**	**0.95**	**0.005**	**0.001**	**-0.04**	**-0.04**
RMSE	**36483**	**35762**	**3389**	**7333**	**36543**	**35766**	**37525**	**36612**
MAE	**26960**	**26816**	**1117**	**2066**	**26954**	**26791**	**26142**	**26074**

The Fig 7 - Fig 9 shows how the selected algorithms are trying to learn the data points. We can clearly see a very good learning nature of the random forest whereas the remaining are failed to learn the patterns.

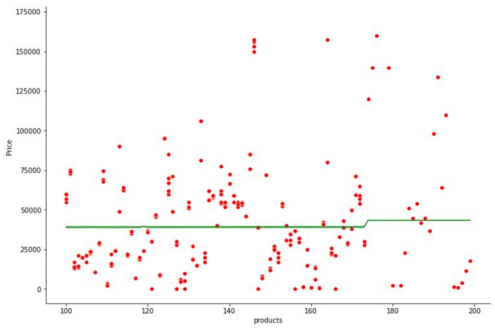


Fig 7 : Lasso Regression

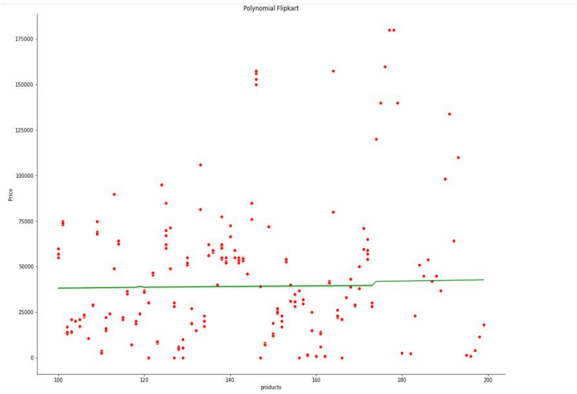


Fig 8: Polynomial Regression

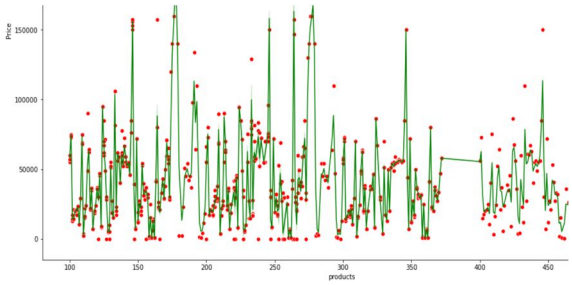


Fig 9: Random Forest Regression

A. Prediction Sample

Table 2 shows the results predicted one days ahead of time. It is observed that the trained random forest model predicted results are almost identical with the actual price on that day. We can see the user can get the best deal of product A in website B and also lose the money if purchased from website D and similarly for product B, user can get the best deal in website C and lose the money if purchased in website D. Users can use the email alert utility. Whenever the product price decreases to the user set price, the user will get an email notification to purchase the product. Fig 10 shows a sample of email notification products.

Table 2 : Proactive Predictions

Product_name	Price(2022-06-17)	Pred_price	Website	Unique_id
Product A	₹20,000	₹19675	Website A	118
Product A	₹18,690	₹19177	Website B	218
Product A	₹20,000	₹20500	Website C	318
Product A	₹23,990	₹25369	Website D	418
Product B	₹19,764	₹19586	Website A	131
Product B	₹26,999	₹27027	Website B	231
Product B	₹16,999	₹1728	Website C	331
Product B	₹22,999	₹2359	Website D	431

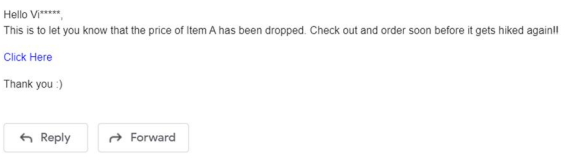


Fig 10: Email alert notification sample

IV. CONCLUSION

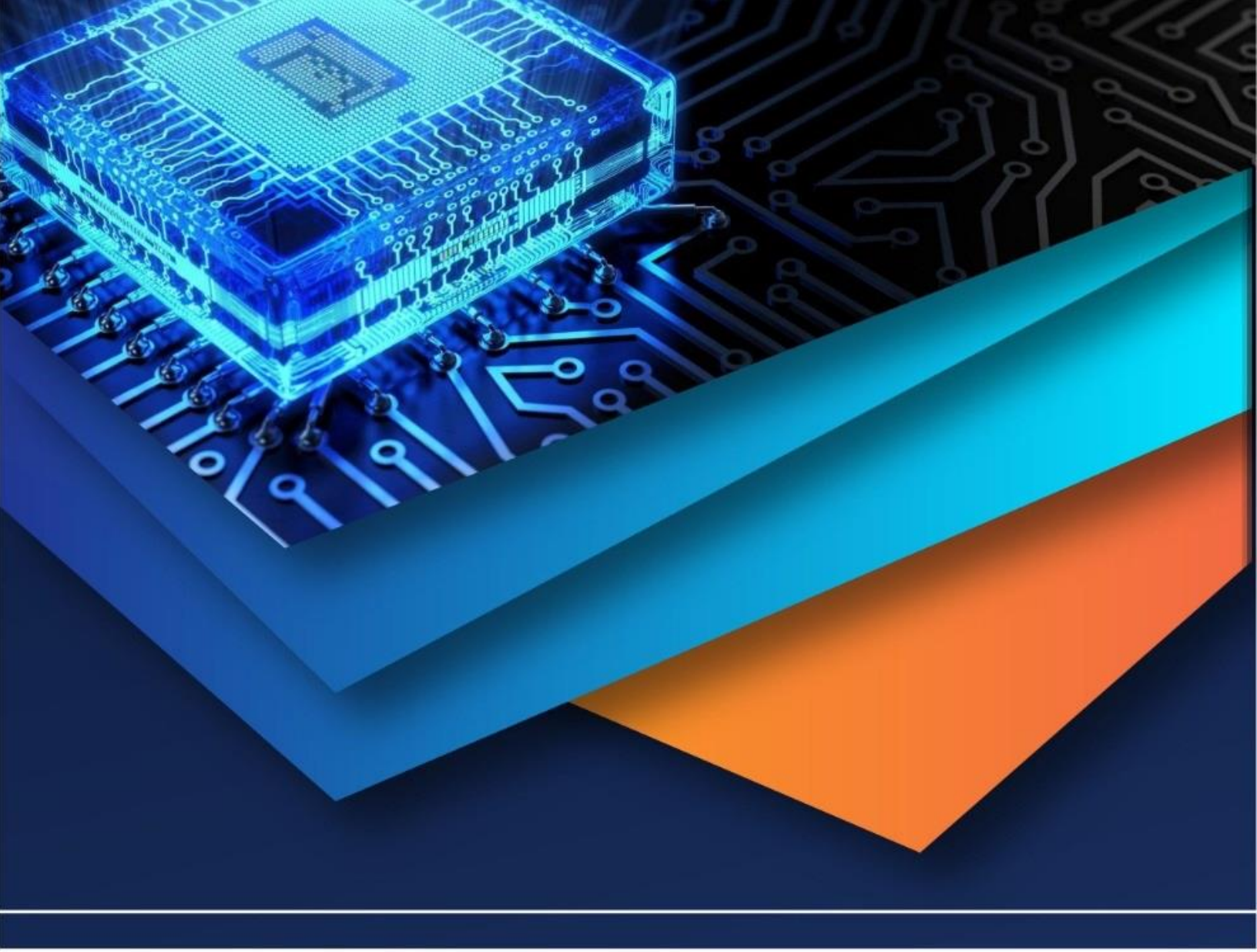
The proposed system is tried on different regression algorithms. Random Forest Regression out performed the Polynomial Regression, Lasso Regression and SVM models. Random Forest Regression showed R-squared of 0.95 and RMSE of 7333 on test data. It is observed that with the increase in days, the R-squared and RMSE metrics improved by learning the patterns of the prices and seasonal effects. The proactive nature of the model benefits the user to estimate the variation in the price of the product and can prepare to buy the product in future. The Email-alert system sends the notification to the user who has set the price limit. As the prices of the products change very frequently, hourly based data scraping gives accurate results from time to time to the users. The proposed system successfully achieves the user expectation while purchasing the products providing the best buy available on E-commerce sites.

Automating the web scraping process, which not only eliminates the manual efforts but also makes the data more time consistent. Developing an UI, web application/mobile application and web extensions by which users can use it easily. Feature exploration and model development for getting better performance results. Implementing more testing models for higher coverage and also eliminating the exceptions, bugs and errors. Making the real time model for prediction and data visualization of the products and their details is always an advantage to the users.

REFERENCES

[1] Singrodia, V., Mitra, A. and Paul, S., 2019, January. A review on web scraping and its applications. In 2019 International Conference on Computer Communication and Informatics (ICCCI) (pp. 1-6). IEEE.
[2] Ullah, H., Ullah, Z., Maqsood, S. and Hafeez, A., 2018. Web Scraper Revealing Trends of Target Products and New Insights in Online Shopping Websites. INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS, 9(6), pp.427-432.
[3] Hillen, Judith. "Web scraping for food price research." British Food Journal (2019).
[4] Milev, Plamen. "Conceptual approach for development of web scraping applications for tracking information." Economic Alternatives 3 (2017): 475-485.
[5] Marques, Pedro, Zayani Dabbabi, Miruna-Mihaela Mironescu, Olivier Thonnard, Alysson Bessani, Frances Buontempo, and Ilir Gashi. "Detecting Malicious Web Scraping Activity: a Study with Diverse Detectors." In 2018 IEEE 23rd Pacific Rim International Symposium on Dependable Computing (PRDC), pp. 269-278. IEEE, 2018.
[6] Bruni, Renato, and Gianpiero Bianchi. "Website categorization: A formal approach and robustness analysis in the case of e-commerce detection." Expert Systems with Applications 142 (2020): 113001.
[7] Kunang, Y.N. and Purnamasari, S.D., 2018, October. Web scraping techniques to collect weather data in South Sumatera. In 2018 International Conference on Electrical Engineering and Computer Science (ICECOS) (pp. 385-390). IEEE.
[8] Scarnò, Marco, and Y. Seid. "Use of artificial intelligence and Web scraping methods to retrieve information from the World Wide Web." Int. J. Eng. Res. Appl. 8, no. 1 (2018): 18-25.
[9] Akrianto, M.I., Hartanto, A.D. and Priadana, A., 2019, November. The Best Parameters to Select Instagram Account for Endorsement using Web Scraping. In 2019 4th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE) (pp. 40-45). IEEE.
[10] Himawan, Arif, Adri Priadana, and Aris Murdiyanto. "Implementation of Web Scraping to Build a Web-Based Instagram Account Data Downloader Application." IJID (International Journal on Informatics for Development) 9, no. 2 (2020): 59-65.
[11] Sundaramoorthy, K., Durga, R. and Nagadarshini, S., 2017, April. Newsone—an aggregation system for news using web scraping method. In 2017 International Conference on Technical Advancements in Computers and Communications (ICTACC) (pp. 136-140). IEEE.
[12] Soujanya, R., Goud, P.A., Bhandwalkar, A. and Kumar, G.A., 2020. Evaluating future stock value asset using machine learning. Materials Today: Proceedings, 33, pp.4808-4813.
[13] Sattarov, O., Jeon, H.S., Oh, R. and Lee, J.D., 2020, November. Forecasting Bitcoin Price Fluctuation by Twitter Sentiment Analysis. In 2020 International Conference on Information Science and Communications Technologies (ICISCT) (pp. 1-4). IEEE.
[14] Vargiu, E. and Urru, M., 2013. Exploiting web scraping in a collaborative filtering-based approach to web advertising. Artif. Intell. Res., 2(1), pp.44-54.
[15] Uzun, E., 2020. A novel web scraping approach using the additional information obtained from web pages. IEEE Access, 8, pp.61726-61740.
[16] Dewi, L.C. and Chandra, A., 2019. Social media web scraping using social media developers api and regex. Procedia Computer Science, 157, pp.444-449.
[17] Nurcahyawati, V. and Mustaffa, Z., 2020, December. Online Media as a Price Monitor: Text Analysis using Text Extraction Technique and Jaro-Winkler Similarity Algorithm. In 2020 Emerging Technology in Computing, Communication and Electronics (ETCCE) (pp. 1-6). IEEE
[18] Pratiba, D., Abhay, M.S., Dua, A., Shanbhag, G.K., Bhandari, N. and SINGH, U., 2018, December. Web Scraping And Data Acquisition Using Google Scholar. In 2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS) (pp. 277-281). IEEE.
[19] Julian, L.R. and Natalia, F., 2015, November. The use of web scraping in computer parts and assembly price comparison. In 2015 3rd International Conference on New Media (CONMEDIA) (pp. 1-6). IEEE.
[20] Alam, A., Anjum, A.A., Tasin, F.S., Reyad, M.R., Sinthee, S.A. and Hossain, N., 2020, June. Upoma: A Dynamic Online Price Comparison Tool for Bangladeshi E-commerce Websites. In 2020 IEEE Region 10 Symposium (TENSYMP) (pp. 194-197). IEEE.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)