International Conference on Recent Trends in Data Science and its Applications
DOI: rp-9788770040723.019

Company Permissibility for Business Activities using NLP

Praseetha P
Computer Science and Engineering
Rajalakshmi Engineering College
Chennai, India
praseetha.p.2019.cse@rajalakshmi.edu. In

Deepthi P
Computer Science and Engineering
Rajalakshmi Engineering College
Chennai, India
deepthi.p.2019.cse@rajalakshmi.edu.in

Dr N Srinivasan
Computer Science and Engineering
Rajalakshmi Engineering College
Chennai, India
srinivasan.n@rajalakshmi.edu.in

***Abstract*— As the amount of web pages available on the visible net has increased to billions of pages with trillions of pages available from the invisible net, extracting data from the internet has become a highly essential procedure in the last few years. Several tools and protocols are made available which can extract this information, and these tools have come in great demand in recent times as researchers and surfers want to discover new knowledge at an ever-increasing rate. Investing in a company requires a lot of analysis to be done before plunging in. As part of this process, the analyst manually does a search of a company name in google, reads search results to determine if there are articles negatively impacting the company and based on this research recommends the company of investment. With enormous data available on the interest, this becomes a very tedious process. In this project, techniques like web scrapping and sentiment analysis are being applied. Libraries like beautifulsoup4, scrapy are being used for scraping the data. The web scraping process is automated using RPA (Robotic Process Automation) making the web scraping process easy as well as efficient. Another technique used is the sentiment approach that applies text analysis algorithms, natural language processing (NLP), and statistics to interpret customer sentiment classifying the customer opinions into positive, negative, or neutral categories. NLTK (Natural Language Toolkit) is applied to perform sentiment analysis.**

Keywords—Web scraping, RPA, Sentiment Analysis

I. INTRODUCTION

Finding answers to issues through the analysis and interpretation of data is known as data analytics. The procedure entails observing and identifying the issues, addressing the availability of pertinent data, determining the approach that will best assist in fixing the issues, and publishing the findings. The process of organizing, cleaning, reanalyzing, using models and algorithms, and eventually producing the end product is how the data is often divided. Data analytics eliminates hunches and manual effort. Companies can use the insights uncovered by data analytics to guide their decision-making and permits you to adjust client care in accordance with their needs. It also provides personalization and improves relationships with customers. Data analysis can reveal information about the preferences, problems, and more of the clientele. It gives you the opportunity to recommend improved products and services.

The objective of this paper is to extract information from different web pages resulting from a web search. The programming language which will be used here is Python. Python libraries are used to scrape the content from each of the links that result from the custom search. RPA bots will be automating the whole web scraping process. The contents arethen organized, analyzed, and cleaned and then on top of itNatural Language Toolkit is applied to perform sentiment analysis. The end result of the analysis would be stored in a csv file and the output is displayed to the user.

II. RELATED WORKS

As there is no specific work-related sentiment analysis for corporate permissibility for business activities utilizing NLP, this section covers information regarding web scraping automation using RPA and sentiment analysis. According to [1]'s authors, text in images can be automatically identified and updated in a target file using text recognition software. The suggested remedy employs a site URL as its information and a web data extraction method to obtain the image or text needed. From a location that the user specifies, the system extracts textual data. Additionally, the retrieved text is classified using Support Vector Machine (SVM) and Naive Bayes Classifier. The output can be saved as an Excel file, CSV file, PDF file, text file, or Google Sheet depending on the user's preferences. UiPath is a brand-new tool that was just recently released to the workplace and is a well-known[3] RPA. This tool functions as a bot that executes predetermined tasks that the user has defined for it by following the directions of a programmed flowchart. This article examines the use of RPA to create learning materials and documents that are used by lots of participants. An RPA was used in the study to test situations involving repeated tasks. It was suggested that automated work management and AI would produce a workflow that was more effective and efficient while lowering the error rate. Web scraping [4] is a technique used to collect data from a website and save the data in files or databases in an organized fashion. The tedious procedure of accessing lists of websites on a regular basis to seek for and store data is automated via web scraping. Manually copying and pasting data into files is a time-consuming and laborious task. An automated web scraping program completes the same task faster. Software for web scraping can be set up to work with any websites or it can be specifically built for a particular website. This study has established a straightforward methodology for analyzing job portals and web scraping information from job descriptions to analyze the needs of the Indian IT sector. [5] the concept of "web design scraping," that basically promotes the idea of extracting, understanding, and modelling website elements and features and so determining the significance of the web design. [6] The method suggested in this study gathers all recipe data using web scraping, after which Python and MongoDB are used to look for recipes that include the specified ingredient. The Python scrappy function is used in online scraping to take the content of a website. Web

scraping methods are used to extract relevant information from websites by scanning hypertext elements and gathering the plain text that is encoded in them fromenormous volumes of web data.

Whether you work as a data analyst, developer, or in another profession that requires you to analyze huge datasets, having the ability to scrape data from websites is a vital talent to have. [7] Through cross- examination and information translation, information examination is a technique for locating solutions to problems. Web data scraping and freely supporting are outstanding techniques for regularly creating content on the web. The computer programs that write web scrubbers are set up to thoroughly mine each key piece of information from numerous internet retailers, compile it into the new website, and then publish it. [8] This study describes how web crawler and NLP can be utilized to deliver sophisticated solutions in computer science education. Using online job postings as a tool, they have examined a specific use of data extraction and Analytics to assess factors that may affect CS undergraduates' ability to find employment after graduation.

[10] The paper Web Scraping and application areas discuses about all the different types of web scraping techniques over the period of time. It also reveals various web scraping libraries which can be used in different areas of interest. It also specifies areas where web scraping is very minorly explored. It also gives an overview of different types of approaches, categories and tool used in this field. [11] aims to remove information from various sources with the aid of tool named as the web crawler Scrapy using the Python 3.6 programming language. A database is made that gathers all the unorganized data from different sources, analyses it according to its specifications by assembling, organizing, cleaning, re- analyzing, using models and algorithms, and then outputs the appropriate results. [16] In general, this paper refers to web scraping as the process of extracting data from a website. Web mining will make it easy to retrieve data from a website. By enabling data scraping from numerous sources, this tactic will reduce the amount of manual labour required, free up time, and increase the usefulness of the data relevant. The user will find it simpler to extract information from sites, store it for his needs, and use it however they see fit as a result of this. The information that has been scraped can be used to create databases, carry out analysis, and accomplish a variety of other tasks. [2] In this paper, They recommended a sentiment analysis technique based on deep learning for e-commerce goods review data. Using this strategy, comments are categorized as either positive or negative. The text is broken up into phrases, and the word length and word frequency are combined to train the neural network. In order to train an emotion classifier and mine the strong correlation between a feature set and an emotion tag, we employ a convolution neural network. According to the experimental findings, the model is a useful tool for online review analysis since it can accurately categorize sentiment and extract useful product attributes. [6] have put up an alternative concept known as aspect-based sentiment analysis, which more accurately recognizes characteristics and achieves the best classification accuracy. In actuality, the concept was created as a smartphone application to help travelers find the best hotel in the area. A number of real-world data sets were used to analyze the model architecture, and the findings showed that the proposed model was effective for both recognition and classification. [12] This paper's primary goal is to review existing sentiment analysis algorithms for Twitter data and present theoretical comparisons of the state-of-art approaches. A variety of approaches to Sentiment analysis at both the document and sentence level are also expressed.

Various approaches to sentiment analysis for Twitter are described, including supervised, unsupervised, lexical, and hybrid approaches. Finally, the latter's discussions and parallels are highlighted. [13] paper examines various well-known approaches or proposals for Sentiment Analysis. To add new components to the suggested plan, the pros and cons of the methodologies mentioned are examined. The new method employs a machine learning mechanism at the document level and combines verbs, adverbs, and adjectives. Along with adverbs, adjectives, and verbs, other combinations that are taken into account for analysis include adjectives-verbs, verb endings, adverbs-adjectives- verbs, and adjectives-verbs. Standard classifiers like Naive Bayes (NB), Linear Model, and Decision Trees are used to infer and interpret the findings. [14] This paper's major goal is to help scholars find the high sentiment analysis-based research publications. In this study, reference phrases are used to analyze sentiment on scientific papers using a previously created annotated corpus. To clean the data corpus, noise was eliminated from the data using several data normalization algorithms. They developed a system that uses six distinct data mining algorithms, such as Nave-Bayes. Lemmatization, n-grams, tokenization, and stop word removal are a few other feature selection approaches that are utilized to improve the system's accuracy. Their method increased outcomes up to 9% above the baseline system. On this data set, classification was performed using the Neural Network, SVM, LR, DT, KNN, and RF. The system's accuracy is then assessed using several assessment metrics such as F-score and Accuracy score. [15] Attempts have been made in this study to anticipate the direction of the stock market, determine prospective pricing for a company's stock, and make other financial judgements service this demand, streaming data appears to be a permanent supply of real-time data analysis. Spark streaming was used for data processing, while data input methods such as Twitter APIand Apache Flume were later used for analysis. The general model for sentiment analysis is done using the Stanford Core NLP tool, which aids in correctly identifying the sentiment of each tweet into three distinct classes: positive, negative, and neutral. [17] This study describes the creation and incorporation of a text analysis workflow into the open public cross-media analysis system. They also compared two major kinds of sentiment analysis methodologies prior to integration, namely lexicon-based and machine-learning approaches. They used the Stanford core NLP library, the Recursive Neural Tensor Network (RNTN) model, the Tomcat architecture, and its lexicon-based sentiment prediction method. Overall, RNTN outperforms the linguistic approach in terms of accuracy for favorable, unfavorable, and neutrality comments of varying length by 9.88%. However, when it comes to classifying compliments, the lexicon-based method works better. We also discovered that F1-score values are 0.16 higher than the

DOI: rp-9788770040723.019

RNTN.[18] In this research, In order to evaluate phrase sentiment analysis for sadness assessment utilizing SVM, NB, and ME classifiers, the author used a vote method and feature selection technique. Two datasets, the Twitter dataset and the20 newsgroups dataset, were used to test the proposed approaches. It shows that SVM outperforms Nave Bayes and Maximum Entropy classifiers in terms of performance. SVM accuracy is 91%, Nave base accuracy is 83%, and Maximum Entropy accuracy is 80%. [19] The objective of this paper is to detect hate remarks on Twitter by categorizingthem as racist, sexist, or neither. They experiment with a verity of classifiers in this research, including Logistic Regression, Random Forest, SVMs, Gradient Boosted decision Trees (GBDTs), and Deep Neural Networks (DNNs). These classifiers feature spaces are defined in turn by task-specific embeddings trained using three deep learning architectures: Fast Text, Convolutional Neural Networks (CNNs), and Long Short-Term Memory Networks (LSTMs). Combinations of CNN, LSTM, and Fast Text embeddings as GBDT features did not produce superior results. Finally, it was discovered that this strategy greatly outperforms the existing methods. The best accuracy values were found if deep neural network model were combined with gradient enhanced decision trees. [20] In order to find the most pertinent and top-notch YouTube videos, this paper proposes a sentiment analysis methodology based on NLP that is applied to user comments. The suggested process involves four steps. In order to set up the subsequent procedure, the comment collection and preparation module first does some linguistic preprocessing on the data (comments) from a specific YouTube movie. Second, NLP-based algorithms are applied to the text after it has been processed in order to create data sets. The positivity and negativity ratings are then computed for the data sets using the sentiment classifier (Sentistrength). The rating was then calculated using the Standard Deviation.

III. Proposed Framework

In this paper, we will suggest an application that will produce the necessary firm facts for the user. Rather than searching for details of a firm website by website, this saves time and yields more efficient results. To carry out the process, we employ tools such as web scraping and sentiment analysis. Web scraping is a software technology that extracts data from web pages. Such software applications typically replicate human internet browsing activities by implementing a low-level Hypertext Transfer Protocol (HTTP) or embedding a full- fledged web browser, such as Internet Explorer or Google Chrome. Web scraping is closely related to web indexing, a practice that most search engines use to index content on the web utilizing web crawlers. Contrarily, web mining focuses primarily on organizing unorganized online data, frequently in Html template, so that it can be saved for subsequent study in a local database or any other type of file structure.

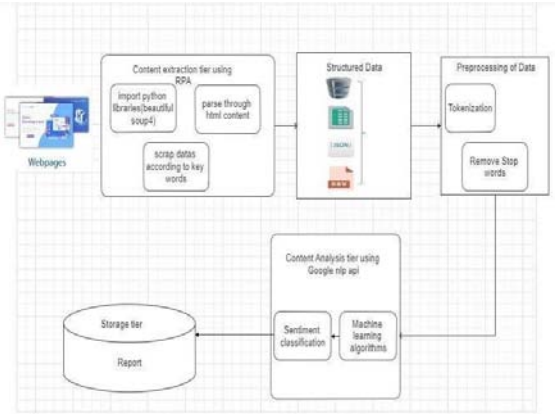

Fig. 1. Architecture Diagram

Robotic process automation (RPA) bots that carry out repetitive operations have automated the entire process. You no longer need to write code each time you gather fresh data from fresh sources. Web scraping is typically made easier andfaster by the RPA systems' built-in features, which also save time. Customer sentiment is interpreted using sentiment analysis, a machine learning technique that combines text analysis algorithms, natural language processing (NLP), and statistics to categorize customer attitudes into positive, negative, and neutral categories. After the content has been organized, analyzed, and cleaned, sentiment analysis will be done using the Natural Language Toolkit (NLTK) to get the sentiment score.

Fig. 1. depicts the architecture of the System. First, the contents from the web pages are scraped in the content extraction tier and the data is stored in the csv file and pre-processing and sentiment analysis are taken place. Then it is displayed in user interface.

DOI: rp-9788770040723.019

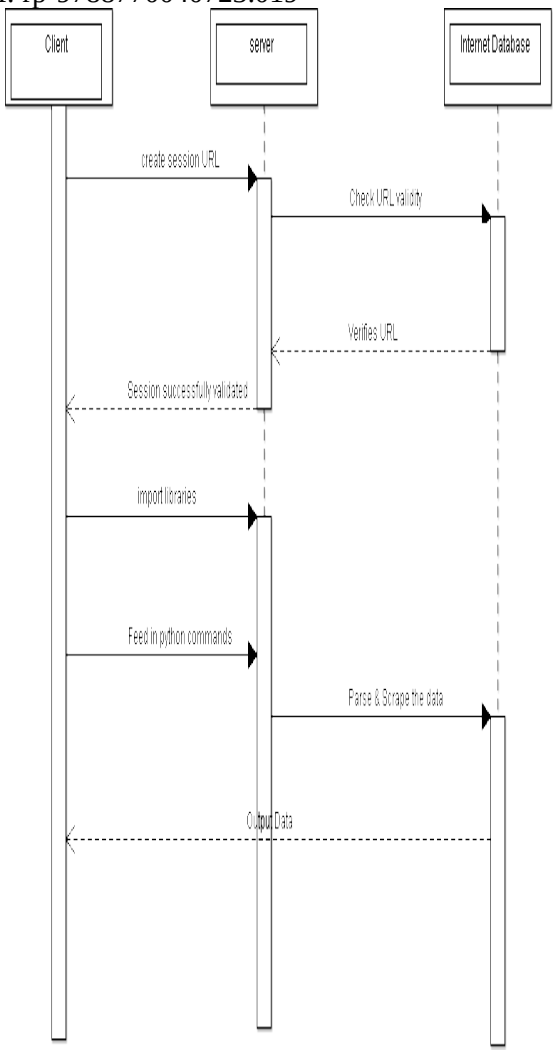

Fig. 2. Sequence diagram for Web Scraping

Fig. 2. depicts the sequence diagram of web scraping where the client creates session url to the server and the server checks the url with the internal database and sends the response back to the client. The interaction between client and internal database is intermediated by the server.

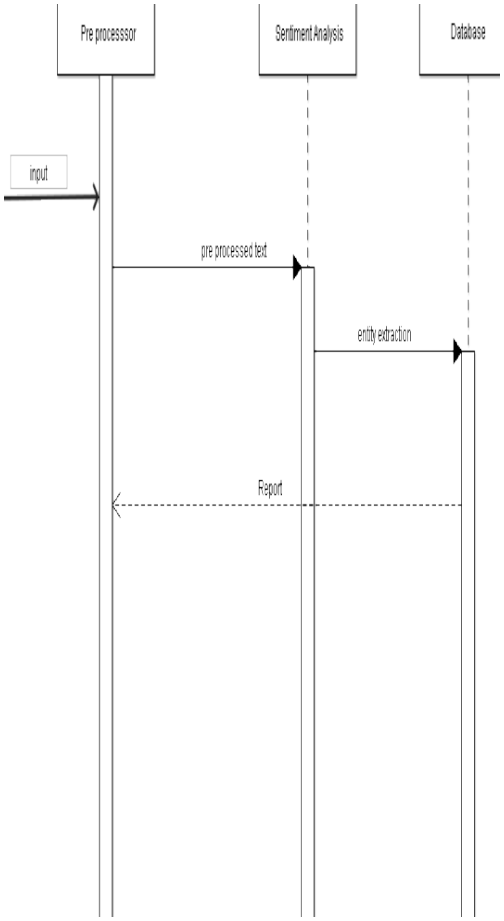

Fig. 3. Sequence diagram for Sentiment analysis

Fig. 3. depicts the sequence diagram of sentiment analysis where the input is given to the preprocessor and the preprocessor cleans and analyses the data and sends it to sentiment analysis tier and result is return to the user. The interaction between client and different tiers are shown in this diagram.

IV. METHODOLOGY

In this work, we'll provide a program that will provide the user with the necessary information about a corporation. Instead of scanning through every website in a company's website directory, this primarily saves time and produces more effective results. To complete the process, we employ methods like sentiment analysis and web scraping. Web scraping is firmly related to web indexing, which is the practice of utilizing a web crawler to index material on the web and is a method that is widely used bymost search engines. Web scraping, on the other hand, focuses primarily on the conversion of unstructured online data—typically in HTML format— into structured data that may be kept for further study in a local database or any other file structure. This whole process is automated using RPA, with RPA there is no need to write code every time you collect new data from new sources. The RPA platforms usually provide built-in tools for web scraping, which saves time and is much easier to use. Sentiment analysis is used to interpret customer sentiment by classifying the customer opinions into positive, negative, or neutral categories. The contents are then organized, analyzed, cleaned and then on top of it

sentiment is performed. The end result of the analysis would be displayed in the UI. As for this project, it will be focused on the web content scraping and sentiment analysis using python as a programming language and automate the process using RPA.

A. Web Scraping using RPA

A Python package called BeautifulSoup can parse XML and HTML pages. For processed pages, it generates a parse tree that can be used to extract HTML data for web scraping. It provides Pythonic paradigms for iterating, searching, and altering the parse tree on top of an HTML or XML parser. We must supply a document to the BeautifulSoup function Object() in order to parse it. HTML entities are translated to Unicode characters before the page is converted to Unicode. The page is then parsed by Beautiful Soup using the built-in HTML parser, unless you specifically tell it to use an XML parser. The HTML content is transformed into a complex tree of Python objects by Beautiful Soup. Four different types of objects are often used: Comment, Tag, NavigableString, and BeautifulSoup. With a different section for each tag and each string, the prettify() method will convert a Beautiful Soup parse tree into a neatly formatted Unicode string. The goal of prettify() is to help understand the structure of the documents you work with. The whole web scrapping process is automated using RPA (Robotic Process Automation) and Orchestrator is used to schedule the process for specified time. The RPA process is carried out in UiPath.

B. *Sentiment Analysis*

The Natural Language Toolkit, or NLTK for short, is a comprehensive open-source framework for developing applications to handle human language data. It includes with powerful text processing modules for standard Natural Language Processing (NLP) operations including cleaning, parsing, stemming, tagging, tokenization, classification, semantic reasoning, etc. The user-friendly interfaces in NLTK are available for Word2Vec, WordNet, VADER Sentiment Lexicon, as well as other popular corpora and lexical resources. Here, sentiment analysis is performed using NLTK. We used NLTK's SentimentIntensityAnalyzer class and the VADER vocabulary to give each comment in the dataset a sentiment score. The Valence Aware Dictionary and Sentiment Reasoner (VADER), a vocabulary and rule-based sentiment analysis toolkit, focuses on the sentiments present in typical text applications such as online comments, social media postings, and survey replies. NLTKanalyses the text and generates the sentiment score for each review. Here, Sentiment analysis is being performed on all the reviews gives by customers for different companies. As an output, we get the sentiment and score value. In order to express the strength of how negative or good the sentiment is, VADER additionally produces a number score that ranges from negative one (-1) to positive one (+1). This is implemented via the SentimentIntensityAnalyzer class's polarity score method, which is known as the polarity score. A negative sentiment is often indicated by a polarity score between -1 and -0.5. In general, neutral sentiment is indicated by a polarity score of more than -0.5 and less than+0.5. Positive emotion is often indicated by a polarity score in the +0.5 to 1 range.

V. RESULT

Web scraping is performed and after which sentiment analysis is carried out and the company details are displayed to the user as output. As a result, the user enters the company name and fetches the score value which is performed with the help of a natural language processor. Web scrapping can be done with the help of many python libraries like beautifulsoup4, scrappy, selenium etc. These libraries can be imported into Pycharm and scrap data by typing lines of code into it. This method is not very efficient since the program has to be run every time in order to be scrapped. More efficient method is to use UiPath's RPA studio which does not require even a single line of code and can be automated by scheduling the time it needs to be scrapped. There are various approaches to perform sentiment analysis. One such way is by importing the csv file into pycharm and perform the analysis by typing several lines of code. It takes quite a long of time to process the code and the chances of producing error are significantly high. Another method is using UIPath studio which also can perform sentiment analysis but it's a tedious process and takes a longer period to produce the output. We found the Sentiment Analysis using NLTK is hassle free and quick, and also gives an accurate score values.

VI. CONCLUSION AND DISCUSSIONS

Web scraping and sentiment analysis are quite prominent in the technology industry in the recent years. When we combine the idea of automating a search with the help of APIs, extracting the content and then performing some analysis on it, the idea becomes vast with the number of applications being limitless. It reduces the manual work of people browsing through websites page by page. This process when automated can reduce the manual overhead thus saving time and energy. Now that title idea is established for automating the search and performing an analysis on that. It can be extended by introducing a few enhancements such as looking for a specific set of keywords or phrases (KWIC, KWAC, KWOC) that the user is expecting to make a decision based on that. The data which is now stored in a database can be presented to the user as a report or dashboard to show the number of web searches done, their corresponding scores and the types of entities it contained. It can contain a word cloud or a color-coded datatext for better visualization. The search can also be parallelized and done in an asynchronous mode to enable faster processing and results for the user.

REFERENCES

[1] N. Roopesh, M.S. Akarsh, and C. NarendraBabu, "An Optimal Data Entry Method, Using Web Scraping and Text Recognition," International Conference on Information Technology (ICIT), 2021.

[2] Li Rong, Zhou Weibai, and Huang Debo, "Sentiment Analysis of Ecommerce Product Review Data Based on Deep Learning," 2021 IEEE 4th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), 2021.

[3] S. Sutipitakwong, andP. Jamsri, "The Effectiveness of RPA in Fine-tuning Tedious Tasks", 2020 6th International Conference on Engineering, Applied Sciences and Technology (ICEAST), 2020.

[4] P. Pillai, and D. Amin, "Understanding the requirements Of the Indian IT industry using web scrapping,"Procedia Computer Science, vol. 172, pp. 308–313, 2020.

[5] A.Namoun, A.Alshanqiti, E.Chamudi, and M. A. Rahmon, "Web

Design Scraping: Enabling Factors, Opportunities and Research Directions", 2020.

[6] S.Chaudhari,R. Aparna, V. G.Tekkur, G. L.Pavan, andS. R. Karki, "Ingredient/Recipe Algorithm using Web Mining and Web Scraping for Smart Chef," 2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), 2020.

[7] Rajesh, M., &Sitharthan, R. (2022). Introduction to the special section on cyber-physical system for autonomous process control in industry 5.0. Computers and Electrical Engineering, 104, 108481..

[8] Lunn, S., Zhu, J., and Ross, M. (2020). Utilizing Web Scraping and Natural Language Processing to Better Inform Pedagogical Practice. 2020 IEEE Frontiers in Education Conference (FIE).

[9] M. Afzaal, M. Usman and A. Fong, "Tourism Mobile App with Aspect-Based Sentiment Classification Framework for Tourist Reviews," IEEE Transactions on Consumer Electronics, vol. 65, no. 2, pp. 233-237, May 2019.

[10] R. Diouf, E. N. Sarr, O. Sall, B. Birregah, M. Bousso, andS. N. Mbaye, "Web Scraping: State-of-the-Art and Areas of Application," 2019 IEEE International Conference on Big Data (Big Data), 2019.

[11] D. M.Thomas,and S. Mathur, "Data Analysis by Web Scraping using Python," 2019 3rd International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2019.

[12] Gomathy, V., Janarthanan, K., Al-Turjman, F., Sitharthan, R., Rajesh, M., Vengatesan, K., &Reshma, T. P. (2021). Investigating the spread of coronavirus disease via edge-AI and air pollution correlation. ACM Transactions on Internet Technology, 21(4), 1-10.

[13] C.Chauhan,and S. Sehgal, "Sentiment analysis on product reviews," 2017 International Conference on Computing, Communication and Automation (ICCCA), 2017.

[14] Hassan Raza, M. Faizan, AhsanHamza, Ahmed Mushtaq, NaeemAkhtar. Scientific Text Sentiment Analysis using Machine Learning Techniques. International Journal of Advanced Computer Science and Applications (IJACSA), vol. 10, no. 12, 2019.

[15] S.Das, R. K.Behera, M.Kumar,and S. K. Rath, "Real-Time Sentiment Analysis of Twitter Streaming data for Stock Prediction,"Procedia Computer Science, vol. 132, pp. 956–964, 2018.

[16] Sameer Padghan, SatishChigle, and Rahul Handoo, "Web Scraping-Data Extraction Using Java Application and Visual Basics Macros,"vol. XV, issue no. 2, (Special Issue) April-2018, ISSN 2230-7540.

[17] Woldemariam, Yonas, [IEEE 2016 IEEE International Conference on Big Data Analysis (ICBDA) - Hangzhou, China (2017.3.12-2016.3.14)] 2017 IEEE International Conference on Big Data Analysis (ICBDA) - Sentiment analysis in a cross-media analysis framework, 2017.

[18] Hassan, AneesUl; Hussain, Jamil; Hussain, Musarrat; Sadiq, Muhammad; Lee, Sungyoung, [IEEE 2017 International Conference on Information and Communication Technology Convergence (ICTC) - Jeju Island, Korea (South) (2017.10.18- 2017.10.20)] 2017 International Conference on Information and Communication Technology Convergence (ICTC) – Sentimentanalysis of social networking sites (SNS) data using machine learning approach for the measurement of depression, 2017.

[19] PinkeshBadjatiya, Shashank Gupta, Manish Gupta, and VasudevaVarma, "Deep Learning for Hate Speech Detection in Tweets," 26th International Conference on World Wide Web Companion, April 2017.

[20] H.Bhuiyan, J.Ara, R.Bardhan,and M. R. Islam, "Retrieving YouTube video by sentiment analysis on user comment," 2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA), 2017.