Harnessing Flask for Web Scraping and Sentiment Analysis: A Comprehensive Application for News and E-Commerce Reviews

Geeta Hanji
Professor
Dept of E & C E
PDACEK, Kalaburagi

Akhil B. Menon
Student
Dept of E & C E
PDACEK, Kalaburagi

Akshay R.
Student
Dept of E & C E
PDACEK, Kalaburagi

Amit Patil
Student
Dept of E & C E
PDACEK, Kalaburagi

Basavaraj Nandeni
Student
Dept of E & C E
PDACEK, Kalaburagi

ABSTRACT

With the popularity and growing availability of opinion rich sources such as reviews from e-commerce sites, choosing the right product from huge product brands is difficult for the user. In order to enhance the sales and customer satisfaction, most of the sites provide opportunity for the users to write review aspects about the product. These reviews are in text format and increase day by day. It is difficult for the user and manufacturers to understand the likes and dislikes of a customer with regard to the product. In this situation sentiment analysis helps the people to analyse the reviews and come to conclusion whether it is good or bad. Sentiment Analysis which is also known as opinion mining is one of the subsection in Natural Language processing in which it learns about sentiment or subjectivity from reviews. The main purpose of the project is to develop a system to extract the reviews from e-commerce site, extract aspect from the reviews and categorize reviews into positive and negative. In this project, the application developed will accept the URL from the user and then ask for the option to be selected. The user can then select whether he/she wants to scrape the news article or the reviews from the blogs or the e-commerce sites. Then the sentiment analysis can be done to get the sentiment of the writer

Keywords

Web Data Extraction, Web Scraping, Selenium, NLTK, Corpa, TextRank algorithm, Sentiment Analysis.

1. INTRODUCTION

Sentiment means ‘opinion’ which is expressed by the people. Analysis of such opinions is known as ‘Sentimental Analysis’. In the recent years, most of the e-commerce websites allow their users to write their opinion about the products which they bought. Situation where in an organization is interested to know its customers opinion regarding the product that it has manufactured, it is very difficult for organization to keep track of each and every opinions. In this scenario sentiment analysis plays a very important role. Sentiment analysis is identifying the polarity of the text, i.e., it identifies the attitude of the user towards a particular topic. Sentiment analysis greatly helps in knowing a person’s behaviour towards a particular topic. If the data is of small size then it is very easy to extract the useful information, but if the size of data is huge then it is quite difficult to analyse what that data actually intends. So there comes the use of Web data extraction. Web data extraction also known as Web scraping is an automated process of extracting the date from the web sites. An automated approach is needed to extract the data and also identify the sentiments expressed in review text. In the context of website Blog, The text is fully scraped and summarised in this research work with the calculation of sentiments of the summarised text. In the context of E-Commerce website, Users’ reviews not only include paragraph of text, but also contain the user and star ratings, quotes, number of people who agree to the particular review and some sentences that are not related to the domain. Manufacturers via the internet get feedback about products through opinions to improve their products. On the other side, customers can read these reviews and decide to buy a product. Reading a large number of reviews is difficult and time-consuming, therefore, it is useful to employ techniques to summarize them automatically. Reasons for Sentiment analysis are message filtering, achieving output in a satisfactory time, and summarizing huge amounts of data. Opinion mining exploits natural language processing and information retrieval techniques to analyze sentiments. The approaches used for sentiment analysis are lexicon based and machine learning based. Lexicon based techniques have low accuracy but they are domain independent. Machine learning classifiers require a lot of training time but they have high accuracy. Therefore, the hybrid method is domain-independent and enhances accuracy.

1.1 Problem Statement

Trust is one of the most important factors in the success of companies, but it is not always easy to establish. It requires businesses to understand their customers and act accordingly. Thanks to emerging technologies such as sentimental analysis, businesses can differentiate and categorize users’ sentiment about their brand or products and services. Since there is a lack of such web applications, we hereby developed the same for News articles and other major E-Commerce sites like Amazon and Flipkart.

1.2 Objectives

The main objective of the work being made is to develop a web application which would be having the capabilities to scrape the data, analyze the product reviews and also understand the sentiment of the writer or user. As per the prepared web application, the project's final interface has the option asking whether to analyze the news article or to analyze the product review from a particular website. Upon selecting the desired option and entering the required url, the output page shows the sentiment of writer. Hence by this way, even the business enterprise may understand the overall perception about their product in the market

2. LITERATURE REVIEW

In the literature, several types of methodologies to perform sentiment analysis and web scraping were found. Firstly for web scraping, the paper proposed by SCM de S Sirisuriya [1], the paper explains neatly the overall process of web data extraction and also explains the different web scraping techniques. The paper also discusses the softwares involved in the web scraping. The paper has a disadvantage that it does not tell about the capacity of any of the techniques thus discussed. The accuracy is also not justified. A research paper published by Vlad Krotov[2] talks about the legality and ethics of web scraping. Next for sentiment analysis a paper by R Raja Subramanian et al[3], defines everything in the context of sentiment analysis, from defining the sentiment analysis, to algorithms for sentiment analysis and from the first step of sentiment analysis to evaluating the predictions of classifiers, additional feature extractions to boost performance are discussed with practical results is given. In the paper authored by Aashutosh Bhatt et al[4] propose a system that performs the classification of customer reviews followed by finding sentiment of the reviews. A rule based extraction of product feature sentiment is also done. The accuracy obtained is not verified in the paper. The sentiment also is changed based on the stars given. The limitations applicable on different sentiment levels were presented by Ansari Fatima Anees et al[5] and those are Fake Inputs, Emotion Detection etc, but at the same time the paper published by Dilesh Tanna et al[6] presents that by using the Algorithms like VADER, we can overcome the limitation of emotion detection. Anisha P Rodrigues et al[7], performed Sentiment analysis with unsupervised machine learning technique like uni-gram Lexicon, bi-gram Lexicon and Supervised technique like Support vector machine (SVM). Although a decent amount of work has been done in the field of web scraping and sentiment analysis independently, a need of integration of both technologies can be found as a future scope of development, hence this is a research paper which talks about the work made in that area.

3. METHODOLOGY

A web application has been developed which has the capability to scrape the data, summarize the news articles, analyze the product reviews and also understand the sentiment of the writer or user.

3.1 Web Scraping

Web Scraping is an automated technique to get the data from the internet. In this research work, Newspaer3k library scrapes the data from the Web articles and the Blogs. Selenium a web browser automation tool is being set up to scrape the reviews from Amazon/Flipkart.

3.2 Sentiment Analysis

The steps involved in sentiment analysis are: Data collection, Preprocessing of data, calculating the sentiment score, sentiment classification of sentiments , evaluation of results.

3.2.1 Data Collection

The first step involved in sentiment analysis is Data Collection. Every analytic task starts with the collection of the data. Nowadays many social media platforms provide an open and easy way to access the data for research and analytics. Real time twitter tweets can be extracted for analysis with twitter developer account and tweepy library in python. Amazon reviews of products can be extracted by web scrape techniques using beautiful soup (also called bs4) library in python. Since bs4 library is not the only method, the research would employ a webpage automation tool known as Selenium to extract the data from E-Commerce website & Newspaper3k for extracting full data from the webpage as it is faster than bs4.

3.2.2 Preprocessing of data

After collecting the data the data needs to be pre-processed. A proper pre-processing shows a big impact on the performance of sentiment classification. The pre-processing can be done differently for different sentiment classification techniques. Pre-processing further involves:

- Tokenisation: It is a process of splitting or dividing the paragraphs into sentences and sentences into words. The words will be stored inside a list named words. Tokenisation in this project is done by the split function.
- Negation handling: In this step, a new set of negative words are being created. The words afters such words are appended with 'not' & such negative words are further removed.
- Stop words removal: Stop words are the words that are not considered or do not contribute in the analysis process. Rather than storing the stop-words in the datasets, it is better to remove the stop-words. So stop words are removed in this process. A new corpa is created named english_stopwords containing all the stop words and those stop words are removed from the text.

3.2.3 Calculating the Sentiment Score

After the removal of stop words, only the words stating the sentiment are left. A corpa named corpa.txt in the form of dictionary is created. According to the words, the sentiment are assigned and all the values are added to get the sentiment score. The total sentiment score is divided by the number of words so that the final sentiment score will be in between -1 to 1.

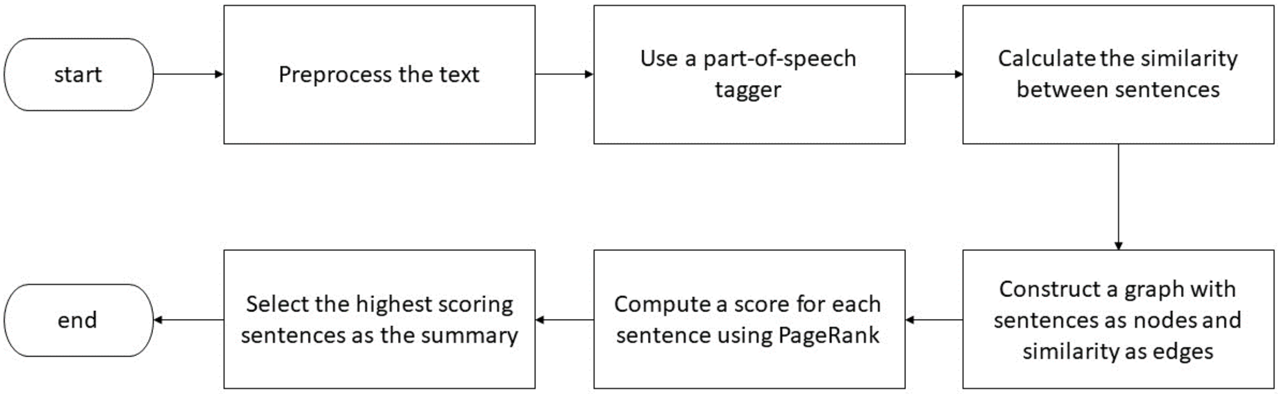
3.2.4 Sentiment Classification

Based on the sentiment scores, the final classification of sentiment is being done. The sentiment score between -1 to 0 is assigned negative, the sentiment score 0 is assigned neutral & the sentiment score between 0 to 1 is assigned positive.

3.2.5 Evaluating Results

After creation of the model the various new words in the real world are checked. The newer words are further added into the corpa.txt, stop words and the negation words so as to improve the accuracy of the algorithm.

Fig 1: Flow diagram of Sentiment Analysis


start
Preprocess the text
Use a part-of-speech tagger
Calculate the similarity between sentences
Construct a graph with sentences as nodes and similarity as edges
Compute a score for each sentence using PageRank
Select the highest scoring sentences as the summary
end

The Fig 1 shows the flow diagram for sentiment analysis. The flow would be understood as below.

- The data which is scraped is taken as input.
- The scraped data is tokenized .
- The tokenized data is handled to check the negativity.
- The stop words are removed
- The sentiment score is obtained by summing up the individual word values from corpa.txt.

3.3 Text Summarization

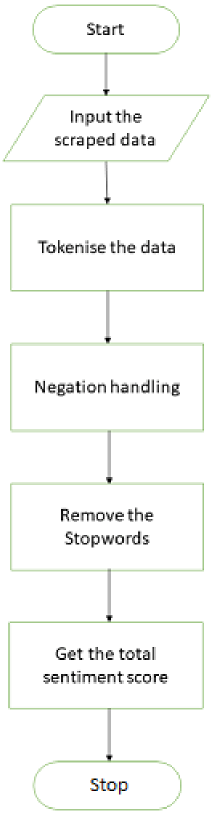


Fig 2: TextRank algorithm

The TextRank algorithm is a graph-based ranking algorithm that was originally proposed for automatic text summarization. The algorithm is based on the PageRank algorithm, which is used by Google to rank web pages in its search engine. The basic idea behind TextRank is to represent a text document as a graph, where each sentence is a node in the graph and the edges between the nodes represent the relationships between the sentences. The Fig 2 shows the implementation of TextRank algorithm.

To build the graph, the TextRank algorithm first preprocesses the text by removing stop words and other irrelevant words, and then uses a part-of-speech tagger to identify the grammatical relationships between the words in the text. The algorithm then constructs a graph where each sentence is a node, and the edges between the nodes represent the similarity between the sentences.

To calculate the similarity between two sentences, the algorithm uses a cosine similarity measure based on the words that appear in the sentences. The algorithm then uses an iterative process to compute a score for each sentence, which is based on the scores of the sentences that are connected to it in the graph.

The final step of the TextRank algorithm is to select the highest scoring sentences as the summary of the document. This can be done by simply selecting the top n sentences, where n is the desired length of the summary.

3.4 Working of the project

The project is a Flask web application that allows users to perform sentiment analysis on website blogs, Amazon product reviews, and Flipkart product reviews. Fig 3, shows the flowchart of the project.

The flow of the project is as below:

1. User Interaction:

- The user interacts with the application through the web interface.
- The initial page displayed is 'index.html', which contains a search bar and three radio buttons for different options.
- The user selects one of the options and enters a URL in the search bar.

2. Option 1: Website Blog Sentiment Analysis:

- If the user selects the first option (Website blog) and submits the form:
- The Flask application receives a POST request to the '/sentiment_or_reviews' route.

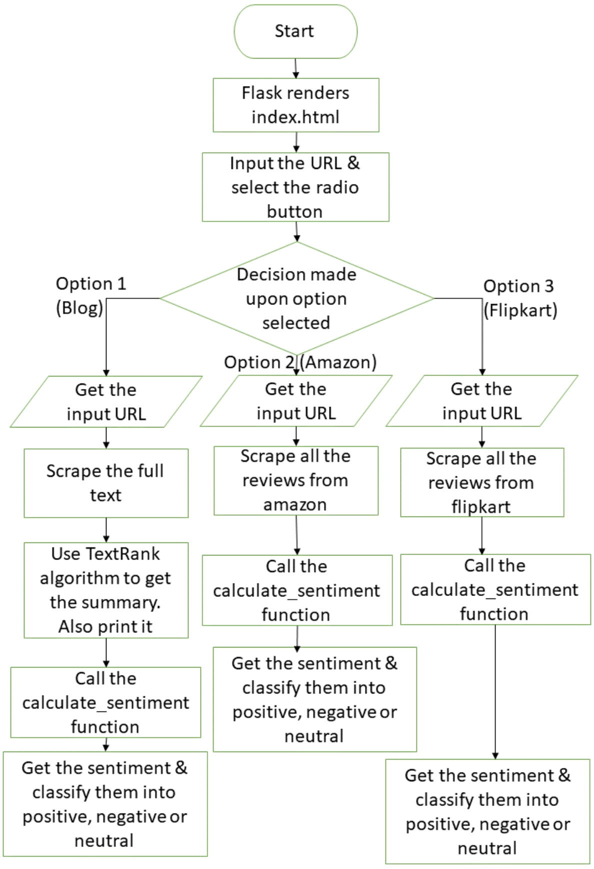

Start
Flask renders index.html
Input the URL & select the radio button
Decision made upon option selected
Option 1 (Blog)
Option 2 (Amazon)
Option 3 (Flipkart)
Get the input URL
Get the input URL
Get the input URL
Scrape the full text
Scrape all the reviews from amazon
Scrape all the reviews from flipkart
Use TextRank algorithm to get the summary. Also print it
Call the calculate_sentiment function
Call the calculate_sentiment function
Call the calculate_sentiment function
Get the sentiment & classify them into positive, negative or neutral
Get the sentiment & classify them into positive, negative or neutral
Get the sentiment & classify them into positive, negative or neutral

Fig 3: Flowchart showing the working of application

- The code downloads the content from the provided URL using the newspaper library.
- The TextRank algorithm is applied to the downloaded content to generate a summary of the text.
- The `calculate_sentiment()` function is called to calculate the sentiment score of the summary.
- The sentiment score is classified as positive, negative, or neutral.
- The original text, summary, sentiment score, and overall sentiment classification are displayed on the 'blog.html' template.

3. Option 2: Amazon Product Review Sentiment Analysis:

- If the user selects the second option (Amazon) and submits the form:
- The Flask application receives a POST request to the '/sentiment_or_reviews' route.
- The code uses Selenium with the Firefox driver to automate browsing.
- The provided Amazon product URL is opened in the browser.
- The code scrolls to the bottom of the page to load all the reviews.
- The reviews are extracted, and for each review:
- The `calculate_sentiment()` function is called to calculate the sentiment score.
- The sentiment score is classified as positive, negative, or neutral.
- The reviews, sentiment scores, and sentiment classifications are displayed on the 'reviews_amazon.html' template.

4. Option 3: Flipkart Product Review Sentiment Analysis:

- If the user selects the third option (Flipkart) and submits the form:
- The Flask application receives a POST request to the '/sentiment_or_reviews' route.
- The code uses Selenium with the Firefox driver to automate browsing.
- The provided Flipkart product URL is opened in the browser.
- The code scrolls to the bottom of the page to load all the reviews.
- The reviews are extracted, and for each review:
- The `calculate_sentiment()` function is called to calculate the sentiment score.
- The sentiment score is classified as positive, negative, or neutral.
- The reviews, sentiment scores, and sentiment classifications are displayed on the 'reviews_flipkart.html' template.

5. Output and Display:

- After performing the sentiment analysis and classification for the selected option, the results are rendered using Flask templates.
- The templates ('blog.html', 'reviews_amazon.html', 'reviews_flipkart.html') display the original text, summary (in the case of a blog), reviews, sentiment scores, and sentiment classifications.

The project uses Flask for the web application framework, newspaper and Selenium for web scraping, and the TextRank algorithm for text summarization. It combines these techniques to analyze sentiment in different contexts (website blogs, Amazon product reviews, Flipkart product reviews) based on user input.

4. RESULTS & CONCLUSION

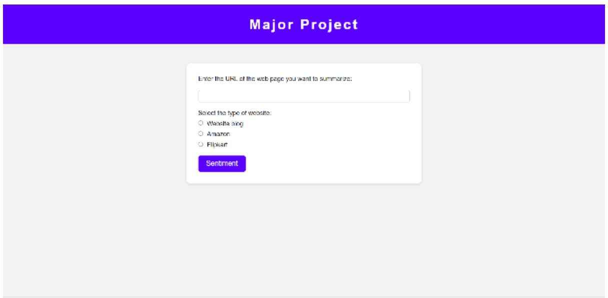
In this project, a web application is developed which would be having the capability to summarize the news article, analyze the product reviews and also understand the sentiment of the writer or user. As per the plan, our project’s final interface will be having the options asking whether to analyze the news article or to analyze the product review from a particular website. Upon selecting the desired option and entering the required

URL, we would get the output showing the sentiment of writer or the sentiment of the product review.

Hence by using this web app, we would be able to differentiate reviews based on their polarity and can address the negative reviews so that the business enterprise may understand the overall perception about their product in the market.

The analysis of the output is being done so as to improve the corpa, stop words and the negation words and hence improve the overall accuracy of the sentiment score being obtained.

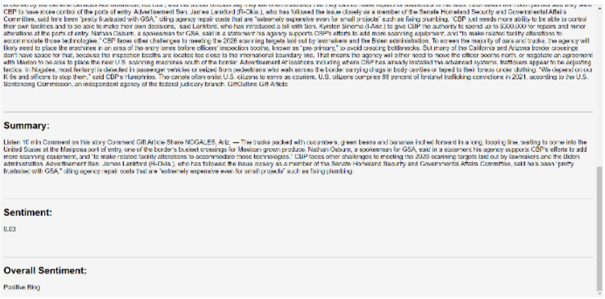
The final outputs from the project can be seen below.



Enter the URL of the web page you want to summarize:

Select the type of website:
○ Website blog
○ Amazon
○ Flipkart

Fig 4: Final interface of the project



CBP to have more control of the ports of entry. Advertisement Sen. James Lankford (R-Okla.), who has followed the issue closely as a member of the Senate Homeland Security and Governmental Affairs Committee, said he's been "pretty frustrated with GSA," citing agency repair costs that are "extremely expensive even for small projects" such as fixing plumbing. "CBP just needs more ability to be able to control their own facilities and to be able to make their own decisions," said Lankford, who has introduced a bill with Sen. Kyrsten Sinema (I-Ariz.) to give CBP the authority to spend up to $300,000 for repairs and minor alterations at the ports of entry. Nathan Osburn, a spokesman for GSA, said in a statement his agency supports CBP's efforts to add more scanning equipment, and "to make related facility alterations to accommodate those technologies." CBP faces other challenges to meeting the 2026 scanning targets laid out by lawmakers and the Biden administration. To screen the majority of cars and trucks, the agency will likely need to place the machines in an area of the entry lanes before officers' inspection booths, known as "pre-primary," to avoid creating bottlenecks. But many of the California and Arizona border crossings don't have space for that, because the inspection booths are located too close to the international boundary line. That means the agency will either need to move the officer booths north, or negotiate an agreement with Mexico to be able to place the new U.S. scanning machines south of the border. Advertisement At locations including where CBP has already installed the advanced systems, traffickers appear to be adjusting tactics. In Nogales, most fentanyl is detected in passenger vehicles or seized from pedestrians who walk across the border carrying drugs in body cavities or taped to their torsos under clothing. "We depend on our K-9s and officers to stop them," said CBP's Humphries. The cartels often enlist U.S. citizens to serve as couriers. U.S. citizens comprise 88 percent of fentanyl trafficking convictions in 2021, according to the U.S. Sentencing Commission, an independent agency of the federal judiciary branch. GiftOutline Gift Article

Summary:

Listen 10 min Comment on this story Comment Gift Article Share NOGALES, Ariz. — The trucks packed with cucumbers, green beans and bananas inched forward in a long, looping line, waiting to come into the United States at the Mariposa port of entry, one of the border's busiest crossings for Mexican-grown produce. Nathan Osburn, a spokesman for GSA, said in a statement his agency supports CBP's efforts to add more scanning equipment, and "to make related facility alterations to accommodate those technologies." CBP faces other challenges to meeting the 2026 scanning targets laid out by lawmakers and the Biden administration. Advertisement Sen. James Lankford (R-Okla.), who has followed the issue closely as a member of the Senate Homeland Security and Governmental Affairs Committee, said he's been "pretty frustrated with GSA," citing agency repair costs that are "extremely expensive even for small projects" such as fixing plumbing.

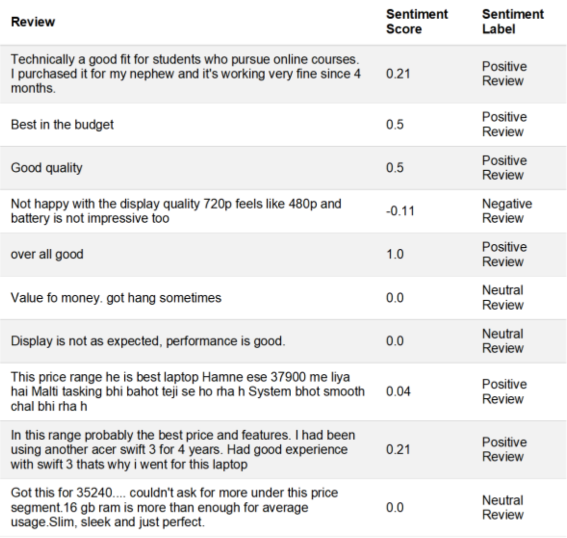
Sentiment:

0.03

Overall Sentiment:

Positive Blog

Fig 5: Summary & Sentiment of the Blog



Review	Sentiment Score	Sentiment Label
Technically a good fit for students who pursue online courses. I purchased it for my nephew and it's working very fine since 4 months.	0.21	Positive Review
Best in the budget	0.5	Positive Review
Good quality	0.5	Positive Review
Not happy with the display quality 720p feels like 480p and battery is not impressive too	-0.11	Negative Review
over all good	1.0	Positive Review
Value fo money. got hang sometimes	0.0	Neutral Review
Display is not as expected, performance is good.	0.0	Neutral Review
This price range he is best laptop Hamne ese 37900 me liya hai Malti tasking bhi bahot teji se ho rha h System bhot smooth chal bhi rha h	0.04	Positive Review
In this range probably the best price and features. I had been using another acer swift 3 for 4 years. Had good experience with swift 3 thats why i went for this laptop	0.21	Positive Review
Got this for 35240.... couldn't ask for more under this price segment.16 gb ram is more than enough for average usage.Slim, sleek and just perfect.	0.0	Neutral Review

Fig 6: Amazon Review page with sentiments

The future scope of the project includes the following advancements

- Improve the algorithm to perform sentiment analysis to increase accuracy.
- May use ML models to improve the overall project.
- May use Deep learning models and transformer based architecture to improve the project.
- Improve the quality of the web application.
- Emotion & Sarcastic detection.
- Can be deployed on cloud so that everybody can access

5. REFERENCES

[1] SCM de S Sirisuriya, "A Comparative Study on Web Scraping", Proceedings of 8th International Research Conference, KDU 2015, pp.135–140.

[2] Vlad Krotov & Leiser Silva, "Web Scraping & it's legality from Legality and Ethics of Web Scraping, Twenty-fourth Americas Conference on Information Systems", New Orleans, 2018.

[3] R Raja Subramanian, Nukula Akshith, Gogula Narasimha Murthy, Manchala Vikas, Srikar Amara, Karnam Balaji, "A Survey on Sentiment Analysis", 2021 11th International Conference on Cloud computing, Data Science & Engineering 2021, pp. 70–75.

[4] Aashutosh Bhatt, Ankit Patel, Harsh Chheda, Kiran Gawande, "Basic Step of Feature Extraction from Amazon Review Classification and Sentiment Analysis", IJCSIT, pp. 5107–5110.

[5] Ansari Fatima Anees, Arsalaan Shaikh, Arbaz Shaikh, Sufiyan, "Survey Paper on Sentiment Analysis: Techniques and Challenges".

[6] Dinesh Tanna, Manasi Dhudhane, Prof. Kiran Deshpande, Amrut Sardar, Prof. Neha Deshmukh, "Sentimental analysis on Social media for emotion detection "

[7] Anisha P Rodrigues, Niranjan N Cahiplunkar, " Aspect based sentiment analysis on product review", IEEE 14th international conference on information processing

[8] Pollyanna Gonçalves, Matheus Araújo, Fabrício Benevenuto, Meeyoung Cha, "Comparing and Combining Sentiment Analysis Methods", pp. 27–37.

[9] Ms. Binju Saju, Ms. Siji Jose, Mr.Amal Antony, "TYPES OF SENTIMENT ANALYSIS from Comprehensive Study on Sentiment Analysis: Types, Approaches, Recent Applications, Tools and APIs", 978-1-7281-6453-3/20/$31.00 ©2020 IEEE.

[10] Scydeh Akram Saadat Neshan, Reza Akbari, " A combination of machine learning and lexicon based techniques for sentiment analysis ", IEEE 2020 6th international conference on web research

[11] Shivaprasad T K, Jothi Shetty, " Sentiment analysis of product reviews: A review ", IEEE International conference on inventive communication and Computational technologies