E-Commerce Web Harvesting using Sentiment Analysis

Mageswari K
Assistant Professor,
Department of Computer science and engineering,
KIT-Kalaignar Karunanidhi Institute of technology.
Coimbatore, India.

Saranya K
Assistant Professor,
Department of Computer science and engineering,
KIT-Kalaignar Karunanidhi Institute of technology.
Coimbatore, India.

Vishalini N
Department of Computer science and engineering,
KIT- Kalaignar Karunanidhi Institute of technology,
Coimbatore

***Abstract*— The number and scope of e-commerce sites have been rapidly increasing with the growth of internet in this modern world making people life easy. Even though the job is easily done people take efforts for finding out the best of all products amonglakhs of products present in it, 92% Consumers look for ratingand reviews regarding the services and products of companies prior to making purchasing decisions. So, here we use web scrapping to find out the products with maximum rating amongtwo most leading e-commerce sites. The scripts are written usingpythonlibraries.Thismakespeople'slifeeasierbygivingthemthebestproduct deals.**

Keywords—webs crapping, python, sentiment analysis

I. Introduction

The use of e-commerce has become very huge during this deadly covid situation. People were brought into a situation of following new norms like social distancing. The announcement of complete lockdown was hard for people to survive and satisfy their daily needs. Even though government and volunteers provided their daily needs, few people preferred shopping by themselves. There are infinite businesses growing up with the help of many upcoming latest technologies, added up with high security and wide spread internet connection in which e-commerce is also one among them. There are many powerful e-commerce sites that allow users to purchase all their needs and post their opinions and satisfaction about the product [8]. Before buying a product it's a practice to check for how good the product is by quantity. Being online shopping, it is impossible to buy the product by looking into such attributes so, the only way to select a product is by viewing the products review which is given by the person who has already purchased the product. Consumers also upload the reviews one-commerce portals but also the upload in other social media networks also, hence feedback plays a vital role [1]. Since now the only option is to do all these activities manually. So, we use web scrapping instead of doing all thesethings manually. One of the advantages of e-commerce is lesstime consuming for purchase and added to this there are many offers put forth. Offers help needy people to get the product at affordable price. Offline shopping requires time and effort tofind the best products whereas online shopping helps shopping anytime from any places the customer is present. The payment procedure takes time in offline shopping in order to manually calculate and produce the bill by the product seller but throughthe modern technology and online transaction process the time for the pay mention lines hopping takes place within seconds.

Since there is huge amount of data present in the e-commerce site, it is hard to search manually. Web scrapping is the technique of data extraction where required data alone are taken from large amount of data [4]. The information is retrieved from the peeks of world wide web. There are data present which are important and required, at the same time there are datapresent in huge quantity which are not in need fat the present. Web scrapping is the process from which we make a semi structured document which are displayed in the format of HTML or XHML through the document from internet, analyzethe document and take only the needy data [9]. Other sites thatdon't allow large amount of data in a structured form or they are simply not that technologically advanced. Sentiment analysis analyses the data and stores them in the text format. Sentiment analysis comes under Natural language Processing that uses Machine Learning algorithms, Lexicon based algorithm and hybrid algorithm. The initial step is to identifythe emotional target pairs then classify the data into highest, lowest, and medium reviews. The analysis mainly targets on aspects terms or feature terms of the product [5]. It is also called as text orientation analysis or opinion mining. It is related toautomata where it is the process of automatically analyzing the subjective commentary text with the customers emotional color and deriving the customers emotional tendency. In recent years, many researches have integrated traditional machine learning methods and deep learning methods into the field of text sentiment analysis by constructing the sentiment lexicon, and achieved good results. RPA is used for comparison between two entities. The major role of RPA technology, to get compared structure of data. We can choose the format in which it should be available to us. This allows ease of access and makes life easier in analyzing the data.

The data extracted and save to the local file in the computer database. Internet is the major source from where we can get amount of data. The user inputs the needed data in the search engine and examines four links to satisfy their data needs. Thereare many ways to gather information from the internet yet web crawling and web scrapping are the two most common ones and while most people use these terms interchangeably, in reality they are not the same thing. Web crawling is the process where the tools are used to read, copy and store the content of the websites for archiving or indexing purposes. Basically, it is one search engine they look through the websites discover what content they include and build entries for search engine index. In Web scrapping large amount of specific data from onlinesources is extracted. The extracted data is further interpreted and parsed by data analyst to make more balanced business decisions. Both are essential methods of collecting data. Data crawling need scrawler bots and scrapping needs scrapper bots. This web extraction is primarily found in enterprises and in the social web sphere. The three methods of implementation of web scrapping are libraries, frameworks, workstation environments. Automatic data extraction reduces the cost, time and manpower.

II.METHOD

This system is brought into the role in order to build a website that produces the result of products which has thehighest rating and also if the same product is put forth withdiscount using web scraping and web scrawling styles. Thewebsite ins designed in such a way that scrapping comes intothe play only when the user input the content that they require. India's two leading e-commerce websites amazon and flipkart are used as the platform from which the scrapping is going totake place. Since internet is the major resource. Generally, theuser types in input they need in the search engine and thatexamines more than a link [3]. Also, in this system Jupyter Notebook plays a big role to write the code which is graphical user interface platform for the user.

WorkingPrinciple

*1.*Importpythonlibraries

Python Requests

In order to scrape a web page, first the required web pagehas to be downloaded. Hence, we use this python request libraryto download the pages. The security is also taken care of the ofthe authentication module. Multiple file sharing is also easilyhandled. The foremost task of the library is to start off with a GET request to the web server. GET request is used to request data from the server. Once they GET request is done it downloads the HTML content in a Web page as it is a humanfriendly HTTP library. There is no necessary to manually add query strings to URLs, or form encode post data. It generally supports browser style support verification, automatic compression, hence from this its major work and support is automation.

b. Be**autifulSoup**

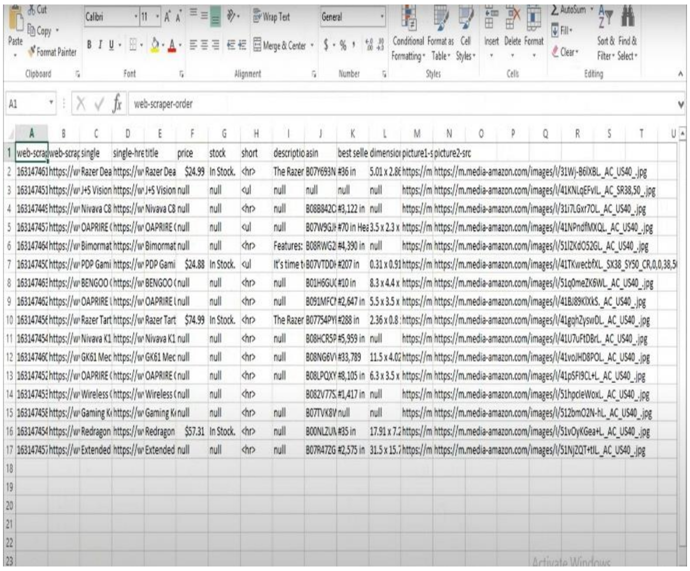
It is a library to parse the document with respect to extract the data of the web page either in HTML or XML format added. Prettify is a method which is used to format the page and make it look better than before. Doc type object which contains information about the type of document. The text in html document is represented through the navigable string. Tag object is the most important object type, which allows to navigate through the HTML document, and extract other tag sand text. The document is converted to UNICODE. The converted Unicode elements become the Unicode characters. Using the children property, the top level of the page is selected. Since all the tags present are in nested from, we can move through the structure one level at a time.

d **Pandas**

Pandas allows to easily read tables within CSV files uploaded to a site from HTML pages. The dataset stored in a CSV file are extracted into DataFrame. It is represented in 2-D data structure. This helps to store and manipulate the tabular data. It shortens the procedure of handling data. The data is filtered according to certain conditions, or segmenting and segregating the data according to preference.

e.**CSVfiles**

As pandas is imported, it is easy to read CSV files available in website. The CSV is a text file that contains data often times the first row of the file is header letting to know what values represent the remaining lines contain the data. Each row has a record in the database. In each row the data areseparated by commas because they are text file and no data types are present. Everything is represented in the form of strings. It is the user responsibility to convert the data into theappropriate data. When there are two commas in a row that simply means the data is missing in the row.



A1 web-scraper-order

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	web-scra	web-scrap	single	single-hre	title	price	stock	short	descriptio	asin	best selle	dimensio	picture1-s	picture2-src
2	163147461	https://w	Razer Dea	https://w	Razer Dea	$24.99	In Stock.	<hr>	The Razer	B07Y693N	#36 in	5.01 x 2.86	https://m	https://m.media-amazon.com/images/I/31Wj-B6lXBL._AC_US40_.jpg
3	163147451	https://w	J+S Vision	https://w	J+S Vision	null	null	<ul	null	null	null	null	https://m	https://m.media-amazon.com/images/I/41KNLqEFvIL._AC_SR38,50_.jpg
4	163147449	https://w	Nivava C8	https://w	Nivava C8	null	null	<hr>	null	B08B842C	#3,122 in	null	https://m	https://m.media-amazon.com/images/I/31i7LGxr7OL._AC_US40_.jpg
5	163147457	https://w	OAPRIRE (	https://w	OAPRIRE (	null	null	<ul	null	B07W9GJH	#70 in Hea	3.5 x 2.3 x	https://m	https://m.media-amazon.com/images/I/41NPndfMXQL._AC_US40_.jpg
6	163147464	https://w	Bimormat	https://w	Bimormat	null	null	<hr>	Features:	B08RWG2	#4,390 in	null	https://m	https://m.media-amazon.com/images/I/51lZXdO52GL._AC_US40_.jpg
7	163147450	https://w	PDP Gami	https://w	PDP Gami	$24.88	In Stock.	<ul	It's time t	B07VTDDH	#207 in	0.31 x 0.91	https://m	https://m.media-amazon.com/images/I/41TKwecbfXL._SX38_SY50_CR,0,0,38,5
8	163147463	https://w	BENGOO (	https://w	BENGOO (	null	null	<hr>	null	B01H6GUC	#10 in	8.3 x 4.4 x	https://m	https://m.media-amazon.com/images/I/51q0meZK6WL._AC_US40_.jpg
9	163147462	https://w	OAPRIRE (	https://w	OAPRIRE (	null	null	<hr>	null	B091MFCN	#2,647 in	5.5 x 3.5 x	https://m	https://m.media-amazon.com/images/I/41BJ89KIXkS._AC_US40_.jpg
10	163147456	https://w	Razer Tart	https://w	Razer Tart	$74.99	In Stock.	<hr>	The Razer	B07754PYI	#288 in	2.36 x 0.8	https://m	https://m.media-amazon.com/images/I/41gqhZyswDL._AC_US40_.jpg
11	163147454	https://w	Nivava K1	https://w	Nivava K1	null	null	<hr>	null	B08HCR5P	#5,959 in	null	https://m	https://m.media-amazon.com/images/I/41U7uFtDBrL._AC_US40_.jpg
12	163147460	https://w	GK61 Mec	https://w	GK61 Mec	null	null	<hr>	null	B08NG6V	#33,789	11.5 x 4.02	https://m	https://m.media-amazon.com/images/I/41voJHD8POL._AC_US40_.jpg
13	163147452	https://w	OAPRIRE (	https://w	OAPRIRE (	null	null	<hr>	null	B08LPQXY	#8,105 in	6.3 x 3.5 x	https://m	https://m.media-amazon.com/images/I/41pSFI9CL+L._AC_US40_.jpg
14	163147453	https://w	Wireless (	https://w	Wireless (	null	null	<hr>		B082V77S	#1,417 in	null	https://m	https://m.media-amazon.com/images/I/51hpcleWoxL._AC_US40_.jpg
15	163147458	https://w	Gaming K	https://w	Gaming K	null	null	<hr>	null	B07TVK8V	null	null	https://m	https://m.media-amazon.com/images/I/512bmO2N-hL._AC_US40_.jpg
16	163147454	https://w	Redragon	https://w	Redragon	$57.31	In Stock.	<hr>	null	B00NLZUN	#35 in	17.91 x 7.2	https://m	https://m.media-amazon.com/images/I/51vOyKGea+L._AC_US40_.jpg
17	163147457	https://w	Extended	https://w	Extended	null	null	<hr>	null	B07R472G	#2,575 in	31.5 x 15.7	https://m	https://m.media-amazon.com/images/I/51NjZQT+tIL._AC_US40_.jpg

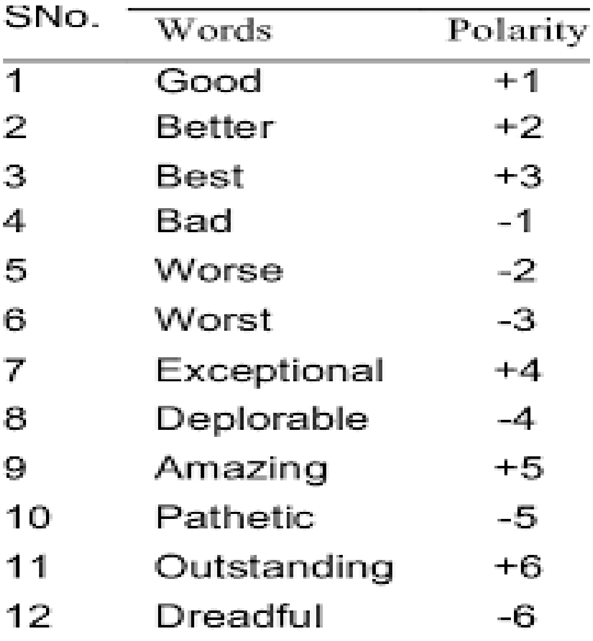
Fig. 1. Amazon CSV file after analysis

III ANALYSIS

Textblob is also one of the python libraries for text processing. Sentiment analysis is a natural language processing technique as it is the process of computationally identifying and categorizing opinions from a piece of text, and determine whether the writer's attitude towards a particular topic or the product is positive, negative or neutral. It might be possible that as an individual we always do not perform sentiment analysis every time but customers do look for the feedback like what the customers has to talk about the product they have purchased that it is good or bad and people analyze manually by looking into these feedbacks. Generally, the e-commerce sites like Flipkart and Amazon in India is a huge hit. The only place it has to touch is the extreme village areas. These e- commerce sites have number of products that Indian people use, the products include from very basic needs to high level gadgets and lots more. Few rare products are not available offline, such products are also made available in these sites. In such situation the person who is new to buy the product for the first time should made to order the product without any hesitation. To tackle such situations, the customer reviews are the helping hand for both the buyer and seller.

The company is also in a need to analyze what the customer thinks about the product but they have more than millions of customers. The very first step is the tokenization. Tokenization is where the data is divided in to statements or dividing the statement into the different set of words. The second step would be the cleaning of the data, by cleaning the data we remove all the special characters or anyother irrelated data which do not add any value to the analytics part. The next step is removing the stop words, stop words donot add any value to the analytics result. Once stop words are removed classification comes into play. The left-out words are classified as positive, negative or neutral word. For a positive word we give a sentiment score as plus one, for negative wegive minus one as the score and for the neutral we give zero as the score.

We can train out model with the bag of words where and test it on analyzing statement. More the accuracy scorebetter will be the classification. If the model is very accurate it is the best classification that has taken according to the previous determined steps. Finally, we combine the statement by adding. Up The scores and perform the needy calculation. Polarity returns positive or negative values. Here positivity refers to the products with highest rating and negativity denotes the products with the least rating [7]. If the polarity is zero then the sentiment of the review is considered to be neural or otherwise positive. As sentiment analysis are basically divided into three stages that is document-stage, sentence -stage, aspect-stage. In sentence level analysis the document is broken into sentences.



SNo.	Words	Polarity
1	Good	+1
2	Better	+2
3	Best	+3
4	Bad	-1
5	Worse	-2
6	Worst	-3
7	Exceptional	+4
8	Deplorable	-4
9	Amazing	+5
10	Pathetic	-5
11	Outstanding	+6
12	Dreadful	-6

Fig. 2. Polarity based on reviews

These sentences are analyzed with the view point of polarity and subjectivity.

Based on the polarity such as negative, positive and neutral reviews as core have been assigned for each specification. By aggregating the scores specific to individual features, on overall product review has been calculated. The success rate of information retrieval depends on the information required and what percentage of overhead the user receives. The Aspect level concentrates more on reviews, feedbacks, comments and complaints.

The existing system of automation involves lot of drawbacks like there are more possibilities of giving false predictions. The first level is document level which is not applicable to documents that evaluate or compare multiple entities, for which fine grained analysis is needed. The sentence level is closely related to subjectivity classification, which decided sentences that expresses factual information that expresses subjective views. Aspect level is based on the idea. Rather than expressing it through positive or negative way, one can also express their view or idea by comparing similar entities, termed as comparative opinions.

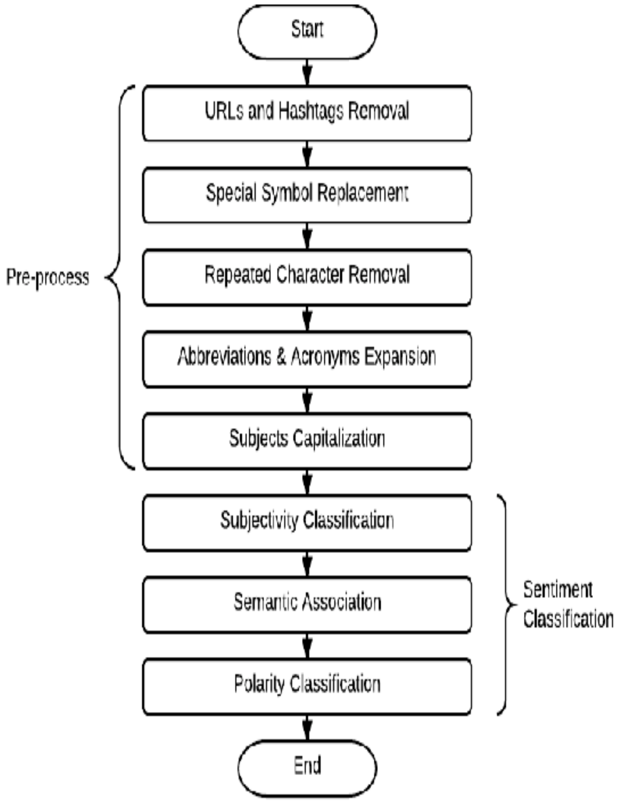
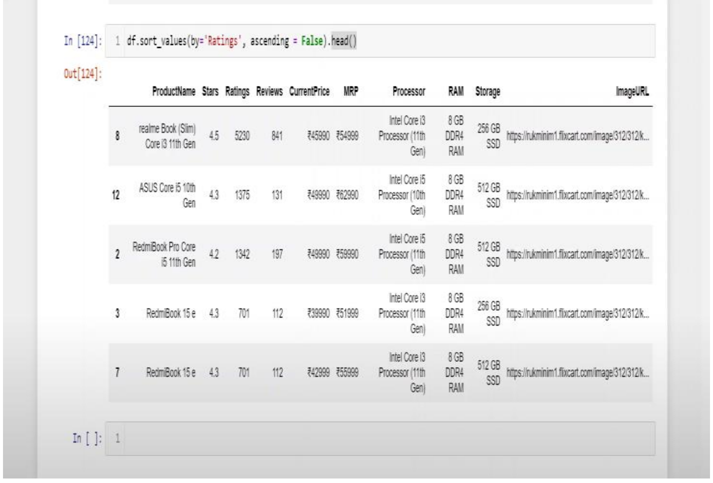

Start
URLs and Hashtags Removal
Special Symbol Replacement
Repeated Character Removal
Abbreviations & Acronyms Expansion
Subjects Capitalization
Pre-process
Subjectivity Classification
Semantic Association
Polarity Classification
Sentiment Classification
End

Fig. 3. Steps in sentiment analysis

This system is proposed to make a user-friendly website for the users such that it acts as the best e-commerce deals for clients from various web domains. Once the customer is satisfied with the product, they further commit to make many purchases in the same e-commerce sites. The customer satisfaction is achieved and also the commerce business sets up a higher mark among every other customer and turns out the offline purchasing to online, which is a new step of turning everything into digitization. The only challenge we are going to face here is negative sentiment using positive words which can be difficult for the machine to detect without having a clear detailed understanding of the context of the situation in which the reviews or the feelings are expressed on the object.



```
In [124]: df.sort_values(by='Ratings', ascending = False).head()
Out[124]:
```

	ProductName	Stars	Ratings	Reviews	CurrentPrice	MRP	Processor	RAM	Storage	ImageURL
8	realme Book (Slim) Core i3 11th Gen	4.5	5230	841	₹45990	₹54999	Intel Core i3 Processor (11th Gen)	8 GB DDR4 RAM	256 GB SSD	https://rukminim1.flixcart.com/image/312/312/k...
12	ASUS Core i5 10th Gen	4.3	1375	131	₹49990	₹62990	Intel Core i5 Processor (10th Gen)	8 GB DDR4 RAM	512 GB SSD	https://rukminim1.flixcart.com/image/312/312/k...
2	RedmiBook Pro Core i5 11th Gen	4.2	1342	197	₹49990	₹59990	Intel Core i5 Processor (11th Gen)	8 GB DDR4 RAM	512 GB SSD	https://rukminim1.flixcart.com/image/312/312/k...
3	RedmiBook 15 e	4.3	701	112	₹39990	₹51999	Intel Core i3 Processor (11th Gen)	8 GB DDR4 RAM	256 GB SSD	https://rukminim1.flixcart.com/image/312/312/k...
7	RedmiBook 15 e	4.3	701	112	₹42999	₹55999	Intel Core i3 Processor (11th Gen)	8 GB DDR4 RAM	512 GB SSD	https://rukminim1.flixcart.com/image/312/312/k...

```
In [ ]:
```

Fig. 4. Products after scrapping process

IV CONCLUSION

Thus, the main goal of the project is to analyze huge datasets and automate so that the customer buys the best products completely without any manual help. This helps the e-commerce sites to help improving the products through the feedback and rating given by the customers. To make this process easier and automate through the best algorithm called sentiment analysis. The input given is the customer review from the best e-commerce sites in India like amazon and flipkart. Before the customer gets the product, they check for the customer review which has already been updated by the customers who have already purchased before. Also, once the product is received the review or rating is given to the same. Here with the concept of web scrapping and sentiment analysis we scrap the data which has only the highest rating. So that the customer easily gets through the best product at the time of purchasing. This acts as an advantage to the customer, product developers, as well as the e-commerce site holders, all the people who are involved in this will achieve a huge success and satisfaction with the product. We analyze the sentiment of aspects with regards to the polarity concept and make more meaningful researches according to the sentiment aspect.

REFERENCES

[1] Bhavna Galhotra “Evolution of E-commerce In India: A Review and Its FutureScope.”,2019.

[2] Singrodia, Vidhi; Mitra, Anirban; Paul, Subrata “A Review on Web ScrappinganditsApplications”,2019.

[3] Ganesh,S,Celestina,A.P.,Jayashree,R.,&HariPriya,K.V.”Web automationinhealthcare”,2019.

[4] Hassan,R.,&Islam,M.R.(2021).Impact of Sentiment Analysis in Fake Online Review Detection.

[5] Yang,Li;Li,Ying; Wang, Jin; Sherratt, R. Simon(2020).“Sentiment Analysis for E-commerce Product Reviews in Chinese based on Sentiment Lexicon and Deep Learning”.

[6] Fan, Xian; Li, Xiaoge; Du, Feihong; Li, Xin; Wei, Mian ,“word vectors for sentiment analysisofAPPreviews.”,2016.

[7] Satuluri Vanaja, Meena Belwal, ”Aspect-Level Sentiment Analysis on E-CommerceData”,2018.

[8] Kaur, Chhinder; Sharma, Anand (2020),”Social issues sentiment analysis using python”,.

[9] Kurniawati, D., &Triawan, D. (2017). “Increased informationretrievalcapabilitiesone-commercewebsitesusingscrapingtechniques”.

[10] Ganesh, S., Celestina, A. P., Jayashree, R., &Hari Priya, K. V.(2019).“Web Automatioin Health Care”.