Big Data Analytics For OTT Platform

M. Jaya Ayiswarya[1], Mr.A. Arul Amalraji[2]
[1]Dept of MCA
[2]Assistant Professor, Dept of MCA
[1, 2] Francis Xavier Engineering College, Vannarpettai, Tirunelveli

Abstract- *The term "Exploratory and Sentiment Analysis" is a conjunction of one after the other specific processes found in the vast field of Data Science. The key to this project is to enhance the value of the Data being utilized, in our case it is Netflix Data – that is an Open-Source Data Set received from Kaggle – that became wrangled and exercised to derive most insights the usage of EDA – Exploratory Data Analysis and Sentiment Analysis after the amalgamation of additional sets – Geographical Latitudes & Longitudes and Netflix Title Critics/Reviews Data Set. The project is made the usage of unique utility analytical tools present in Python Library of flexible programs. This paper introduces systematic and insightful utilization of strategies for Exploratory Data Analysis & Sentiment Analysis through utilising diverse programs concerned.We can say that data visualization is basically a graphical representation of data and information. Day via way of means we're moving closer to information-pushed world. It is relatively useful in order to make choices from information and use the ability of visualization to inform testimonies approximately what, when, where, and the way information would possibly lead us to a fruitful outcome.Data visualization is being used in almost every sector to understand the data better. Because it can be challenging to interpret data from a CSV file, on the other hand, it becomes much easier to understand the data when represented in a chart or map.*

Keywords- Exploratory Data Analysis; Sentiment Analysis; Data Analytics; Python; Seaborn; Numpy;Python library;Juypter notebook;Heatmap)..

I. INTRODUCTION

The term "Data Analysis" is acknowledged to be rooted in the information space, which itself is understood to have an extended history. With the assist of the statistical improvement strategies, we can derive interesting outcomes. The advancement of fast technological implications in the world led to a consequent advent of Big Data; we are constantly being faced with enormous amounts of raw facts which is subject to destiny improvements based on the required parameters and standards through an entity. Starting with the collection of facts, the maximum not unusual place and next step is to carry out the evaluation of it. Data evaluation is for this reason acknowledged to be a systematic manner entirely centered on the facts as its subject. It starts off evolved with retrieving facts from various external-cum-internal sources and then appearing intrinsic evaluation with the facts as a way to find out and obtain beneficial information catering the needs of an entity. For example, the evaluation of population growth through a district can assist governments determine the number of hospitals that would be needed in a given area. When collecting the optimal facts for evaluation, it must hold the minimal viability in phrases of functions and attributes appropriate for our evaluation. This can be represented in phrases of physical and health-orientated functions like Health Status, Age, Male:Female Ratio, BMI etc., will offer a lot more problem precise insights over the population. It can permit someone to visually represent these functions as per the requirements. Fundamentally, there are two primary techniques for facts evaluation – based on the nature and function of facts - qualitative facts evaluation and quantitative fact's evaluation strategies. These facts evaluation strategies have the scope to be utilized independently or in combination with different techniques as a way to benefit get entry to a number of the quality enterprise and intelligence-orientated insights for making higher choices over the already gift facts.

In this paper,systematic and insightful utilization of strategies for Exploratory Data Analysis & Sentiment Analysis through utilising diverse programs concerned. The second section gives the literature review followed by the theory of the method. The fourth section gives simulation environment, experimental results, and performance metrics. The fifth section proceeds with the conclusion followed by thefuture enhancement.

II. RELATED WORK

Kiranbala Nongthombam , Deepika Sharma et al [1], The very basic processes ofdata analysis like cleaning, transforming, modeling of data is briefly explained in this paper and focus more on exploratorydata analysis of an already existing dataset and finding theinsights.

Jyoti Budhwar, Sukhdip Singh et al [2], Data set of review of customer has been considered in order to perform

sentiment analysis. The proposed research is supposed to resolve the issues of previous research that were faced during sentiment analysis

Soniya Grace et al [3],The spatial contour map of these groundwater quality parameters was derived in Arc Map10.5 software using an Inverse Distance Weighted (IDW) spatial interpolation technique. The study facilitates to understand the existing groundwater quality conditions and to develop appropriate management practices to protect the groundwater sources.

Gants J., Reinsel D. et al [4], Advances in the field of technology enabled individuals and businesses to collect large amounts of data (structured and unstructured) from various sources like never before. Data from social media, user-generated, internet, health care, manufacturing, supply chain, financial institution, and sensors have grown exponentially.

Viv Bewick, Liz Cheek, and Jonathan Ball et al [5], The present review introduces methods of analyzing the relationship between two quantitative variables. The calculation and interpretation of the sample product moment correlation coefficient and the linear regression equation are discussed and illustrated. Common misuses of the techniques are considered.

Dr Ossama Embarak, Embarak, and Karkal et al [6], Moving on to data visualization, you will learn how it caters to modern business needs and is key to decision-making. You will also take a look at some popular data visualization libraries in Python.

Nils Gehlenborg and Bang Wong et al [7], Heat maps represent two-dimensional tables of numbers as shades of colors. This is a popular plotting technique in biology, used to depict gene expression and other multivariate data. The dense and intuitive display makes heat maps well-suited for presentation of high-throughput data.

Fabio Nelli et al [8], Python Data Analytics examines how to go about obtaining, processing, storing, managing and analyzing data using the Python programming language.pandas are covered; it is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for Python.It shows the strength of the Python programming language when applied to processing, managing and retrieving information.

Wes McKinney et al [9], A data product, is a computer application that takes datainputs and generates outputs, feeding them back into the environment. It may be based on a model or algorithm.

Matthieu Komorowski, Dominic C Marshall, Justin D Salciccioli, and YvesCrutain et al [10], Training Set and Summary Attribute of Netflix Data as the Input/ Testing Data. As the above EDA catered to our preprocessing needs, where in we extracted the useful features, we have now combined three datasets together to form the training set.

III. THEORY

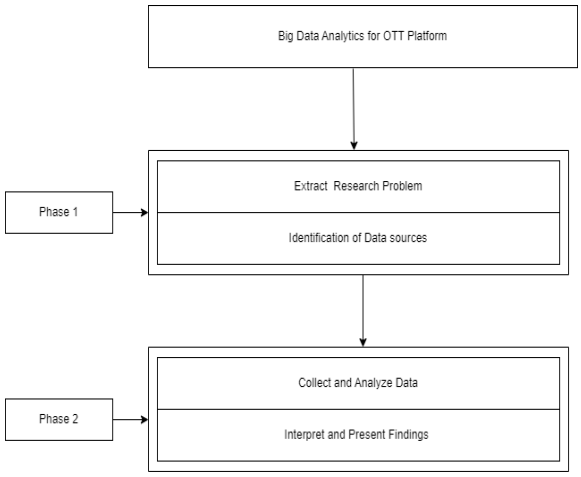
The existing work is that with the advent of technology, the need for consumption of data has increased tremendously. Every single activity of our life has become interlinked with data. As interesting as it may sound, the fact that there is a constant need for intelligent solutions with data as inputs cannot be neglected. Due to these challenges, we have witnessed a paradigm shift in almost every single field around the world. One such field is the entertainment industry that had a rapid transposition as the industry soon began to adopt virtual methods for releasing its content. We can handle huge amount of data at the same time it hardly represents the data in clear visual and it confuses the Overall report.

The proposed work is that Data Visualization focuses on deriving crucial insights from the publicly available Netflix dataset (obtained from Kaggle). It also made use of Geographical Data Set and Netflix Title Reviews Data Set. Once our required datasets are generated – with inculcated properties like optimal, structured and human-understandable data format– we can nowproceed further for an in-depth analysis of it. To begin with, we have chosen EDA as our primary step to analyze our data. We have applied a variety of techniques to gain maximum insights from our data set.Our eyes are drawn to colours and patterns. We can quickly recognize blue from yellow, circle from a square. Data visualization is a form of visual art that not only grabs our interests but also keeps our eyes on the message. And it visualizes the exact data for what we retrieve from dataset and graph that represents a specific category of data with rectangular bars with length and height proportional to the values they represent.

A 1. Research Methodology

When conducting research on big data analytics for an OTT platform like Netflix, it is important to follow a rigorous methodology that ensures the reliability and validity of the results. In this case, the research problem could be something like "How does Netflix use big data analytics to

improve user experience and drive engagement on its platform"to review the existing literature on big data analytics and OTT platforms, including academic research, industry reports, and news articles. This can help to identify gaps in the current knowledge and provide a foundation for the study.



Research Methodology

A 2. Algorithm Implementation

Sentiment analysis is a branch of natural language processing dedicated to the study of subjective opinions and sentiments gathered from various sources. Sentiment Analysis is a set of tools for identifying and extracting sentiment and using it to benefit business operations. Such algorithms dig deep into the text to determine attitudes towards the product in general or its specific elements. In short, opinion mining and sentiment analysis provide an opportunity to explore the mindset of listeners and study the state of a product from opposing perspectives. Here, we using the rule based approach. Rule-based sentiment analysis is based on algorithms that clearly define the opinions to be identified. Includes subjectivity, polarity, or object of opinion. The rule-based approach includes basic natural language processing routines. These are the following operations ona text corpus: stemming, tokenization, part-of-speech tagging, parsing, and lexical analysis (depending on the relevant context). The way it works is that you have two wordlists. One of them contains only positives and the other contains negatives. The algorithm examines the text to find words that match the criteria. Next, the algorithm calculates the parts of speech that occur more frequently in the text. A text is considered of positive polarity if it has more positive words. The problem with rule-based algorithms is that while they produce some kind of result, they lack the flexibility and precision to be useful in practice. For example, the rule-based approach does not take context into account. However, it can be used to set the tone of a message for general purposes. This is useful for customer support. Rule-based sentiment analysis is often used to lay the foundation for subsequent implementation and training of machine learning solutions.

Exploratory data analysis (EDA) is an approach to analyzing data using visual techniques. It is used to discover trends, patterns, or test assumptions using statistical summaries and graphs. Exploratory data analysis (EDA) is used by data scientists to analyze and explore data sets and summarize their key characteristics, often using data visualization techniques. It helps you decide how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, discover anomalies, test hypotheses, or validate assumptions. EDA is primarily used to ensure that data can be revealed beyond the tasks of formal modeling and hypothesis testing, helping to better understand the variables of a dataset and the relationships between them. It also helps determine whether the statistical method you are considering for data analysis is appropriate.

IV. EXPERIMENTS AND RESULTS

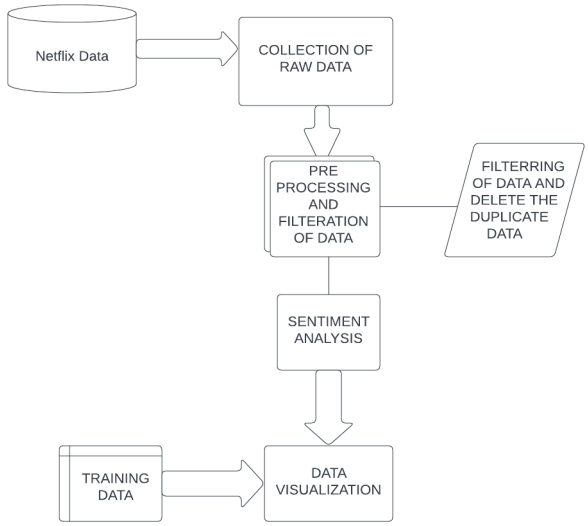
A 1. Simulation Environment

A Jupyter Notebook is an open source web application that allows data scientists to create and share documents that include live code, equations, and other multimedia resources.Jupyter notebooks are used for all sorts of data science tasks such as exploratory data analysis (EDA), data cleaning and transformation, data visualization, statistical modeling, machine learning, and deep learning. Jupyter notebooks are especially useful for "showing the work" that your data team has done through a combination of code, markdown, links, and images. They are easy to use and can be run cell by cell to better understand what the code does.

The Jupyter Notebook is an open-source web application that allows data scientists to create and share documents that integrate live code, equations, computational output, visualizations, and other multimedia resources, along with explanatory text in a single document. You can use Jupyter Notebooks for all sorts of data science tasks including data cleaning and transformation, numerical simulation, exploratory data analysis, data visualization, statistical modeling, machine learning, deep learning, and much more. Jupyter notebooks can also be converted to a number of standard output formats (HTML, Powerpoint, LaTeX, PDF, ReStructuredText, Markdown, Python) via the web interface. This flexibility allows data scientists to easily share their work with others. A Jupyter notebook consists of his two components: a frontend web page and a backend kernel. On

the front-end web page, data scientists can enter programming code or text into rectangular "cells." The browser then passes the code to the backend kernel. The backend kernel executes code and returns results.

A 2. Architecture diagram



Architecture Diagram

These are the types of processing we have done in the in this analysis

Data Requirerments

Data are the most important unit in any study. Data must be provided as inputs to the analysis based on the analysis' requirements. The term "experimental unit" refers to the type of organization that would be used to gather data (e.g., a person or population of people). It is possible to identify and obtain specific population variables (such as height, weight, age, and salary). It doesn't matter whether the data is numerical or categorical.

Data Collecting

The collecting of data is simply known as Data Collecting. Data is gathered from a variety of sources, including relational databases, cloud databases, and other sources, depending on the study' needs. Field sensors, such as traffic cameras, satellites, monitoring systems, and so on, can also be used as data sources.

Data Processing

Data that are collected must be processed or organized for analysis. For instance, these may involve arranging data into rows and columns in a table format (known as structured data) for further analysis, often through the use of spreadsheet or statistical software.

Data Cleaning

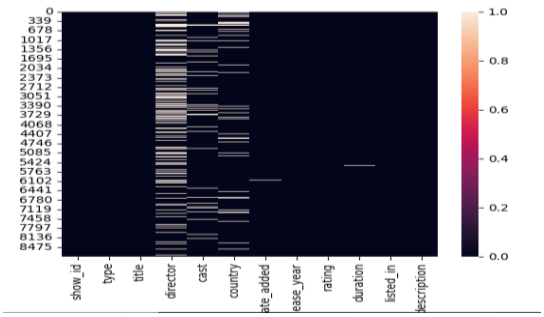
The method of cleaning data after it has been processed and organized is known as data cleaning. It scans for data inconsistencies, duplicates, and errors, and then removes them. The data cleaning process includes tasks such as record matching, identifying data inaccuracy, data sort, outlier data identification, textual data spell checker, and data quality maintenance. As a consequence, it keeps us from having unexpected outcomes and assists us in delivering high-quality data, which is essential for a successful outcome.

Exploratory data analysis

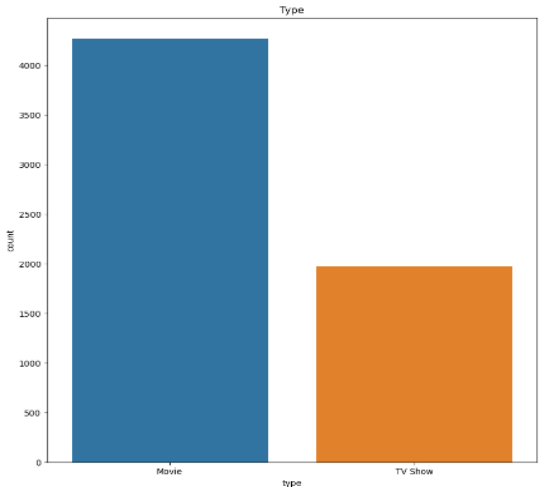
Once the datasets are cleaned and free of error, it can then be analyzed. A variety of techniques can be applied such as exploratory data analysis- understanding the messages contained within the obtained data and descriptive statistics finding average, median, etc. Data visualization is also a technique used, in which the data is represented in a graphical format in order to obtain additional insights, regarding the information within the data

```
  show_id     type                  title           director  \
0      s1    Movie   Dick Johnson Is Dead  Kirsten Johnson
1      s2  TV Show           Blood & Water              NaN
2      s3  TV Show               Ganglands  Julien Leclercq
3      s4  TV Show  Jailbirds New Orleans              NaN
4      s5  TV Show           Kota Factory              NaN

                                                cast        country  \
0                                                NaN  United States
1  Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...   South Africa
2  Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...            NaN
3                                                NaN            NaN
4  Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...          India

          date_added  release_year rating   duration  \
0  September 25, 2021          2020  PG-13     90 min
1  September 24, 2021          2021  TV-MA  2 Seasons
2  September 24, 2021          2021  TV-MA   1 Season
3  September 24, 2021          2021  TV-MA   1 Season
4  September 24, 2021          2021  TV-MA  2 Seasons

                                           listed_in  \
0                                      Documentaries
1    International TV Shows, TV Dramas, TV Mysteries
2  Crime TV Shows, International TV Shows, TV Act...
3                                Docuseries, Reality TV
4  International TV Shows, Romantic TV Shows, TV ...

                                         description
0  As her father nears the end of his life, filmm...
1  After crossing paths at a party, a Cape Town t...
```

Dataset related to Netflix

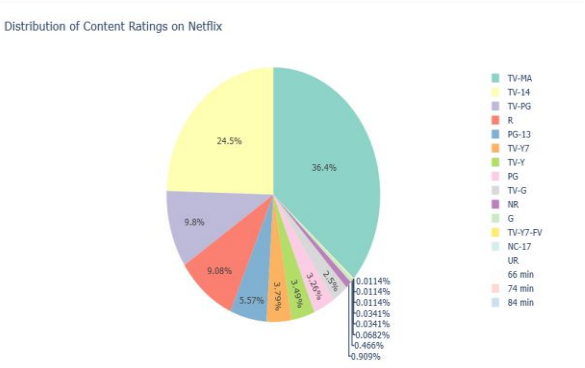


Cleaning the data



Display the movies and shows

Display the categories of movies in wordcloud

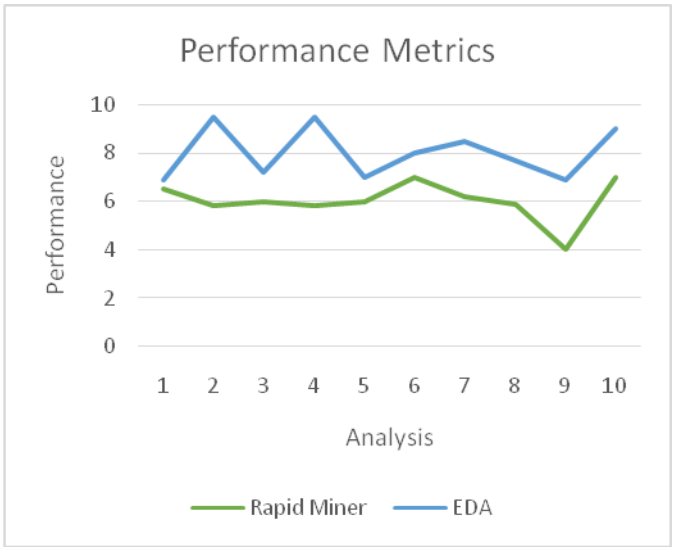


Distribution of content rating on Netflix

A 3. Performance Metrics

Exploratory Data Analysis (EDA) is a crucial step in any data analysis project, including sentiment analysis. EDA helps to gain a better understanding of the data, identify patterns and relationships, and detect potential issues that may affect the analysis.In sentiment analysis, EDA can involve analyzing the distribution of sentiment labels across the dataset, identifying the most frequent words and phrases in positive and negative reviews, exploring the relationship between sentiment and other variables such as time or demographics, and visualizing the data in various ways.

Another tool which is to be compared with EDA is Rapid Miner. RapidMiner is a data mining and machine learning software platform that can be used for sentiment analysis. It provides a range of tools and techniques for data preparation, feature engineering, model building, and evaluation, making it a popular choice for sentiment analysis projects.



V. DISCUSSION AND CONCLUSION

In this paper, various phases of data analysis including data collection, cleaning and analysis are discussed briefly. Explorative data analysis is mainly studied here. For the implementation, Python programming language is used. For detailed research, jupyter notebook is used. Different Python libraries and packages are introduced. Using various analysis and visualization methods, numerous results are extracted. The dataset “Netflix Data Analytics” is used and extract important information’s like the difference in the score of happiness of different countries, the dependence of one attribute in building up the score, how a variable affects another variable, etc. are seen in this analysis and various graphs has been plotted using various attributes in the dataset and draw conclusions in an easy way. Data Analysis is a fundamental step to address the various needs of a client in any professional spectrum. The varied range of insights that can be derived from a data is itself primarily valuable in nature as there are multiple businesses that are actively looking for futuristic, predictive and descriptive insights from the already present raw data generated by them. It helps the organizations to gain access to numerous concealed patterns, information and bits of knowledge after the analysis had been performed. The analysis that we have just performed using the Netflix data not only provides us with incentives to take smart and

intelligent business decisions, but also contribute to the overall growth of the firm.

VI. FUTURE SCOPE

These insights maintain a clear sight and perspective for various stakeholders and help in targeting a positive vision for the future. The future scope of Data Analysis is bound to remain intact as long as businesses require Data Science in their everyday applicable decision-making processes.Also, there is a great scale of possibilities when it comes to developing unique interactive solutions and methods that are confined to make data exploration much more intriguing in nature.These constant advancements have stabilized a promising direction for data analysis as a systemic study that is going to stay as long as there is the crunch for data in any viable field of study in the real-world.

REFERENCES

[1] Kiranbala Nongthombam , Deepika Sharma, 2021, Data Analysis using Python, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 10, Issue 07 (July 2021)

[2] Jyoti Budhwar, Sukhdip Singh, 2021, Sentiment Analysis based Method for Amazon Product Reviews, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) ICACT – 2021 (Volume 09 – Issue 08)

[3] Soniya Grace, 2020, A Geospatial Analysis of Ground Water Quality Mapping using GIS in Sangareddy District, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 09, Issue 07 (July 2020)

[4] Gupta, Bhumika & Negi, Monika & Vishwakarma, Kanika & Rawat, Goldi & Badhani, Priyanka. (2017). Study of Twitter Sentiment Analysis using Machine Learning Algorithms on Python. International Journal of Computer Applications. 165. 29-34. 10.5120/ijca2017914022.

[5] Viv Bewick, Liz Cheek, and Jonathan Ball. Statistics review 7: Correlation and regression. Critical care, 2003.

[6] Dr Ossama Embarak, Embarak, and Karkal. Data analysis and visualization using python. Springer, 2018.

[7] Nils Gehlenborg and Bang Wong. Heat maps. Nature Methods, 2012 Fabio Nelli. Python data analytics: Data analysis and science using PANDAs, Matplotlib and the Python Programming Language. Apress, 2015.

[8] Wes McKinney. Python for data analysis: Data wrangling with Pandas, NumPy, and IPython. " O'Reilly Media, Inc.", 2012

[9] Matthieu Komorowski, Dominic C Marshall, Justin D Salciccioli, and Yves Crutain. Exploratory data analysis. Secondary analysis of electronic health records, 2016.

[10] David F Williamson, Robert A Parker, and Juliette S Kendrick. The box plot: a simple visual method to interpret data. Annals of internal medicine, 1989.

[11] Aashutosh B, Ankit Patel, Harsh Chhedaet al, Amazon Review Classification and Sentiment Analysis / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (6) , 2015, 5107- 5110

[12] Emil Person ,Evaluating tools and techniques for web scraping, Degree Project In Computer Science and Engineering, Second Cycle, 30 Credits Stockholm, Sweden 2019

[13] D. MALI, M. ABHYANKAR, SENTIMENT ANALYSIS OF PRODUCT REVIEWS FOR E-COMMERCE RECOMMENDATION ,International Journal of Management and Applied Science, ISSN: 2394-7926 Volume-2, Issue-1, Jan.-2015,127

[14] Abdullah Alsaeedi1 , Mohammad Zubair Khan, "A Study on Sentiment Analysis Technique of Twitter Data", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 2, 2019

[15] Aliza Sarlan, Chayanit Nadam, Shuib Basri ,"Twitter Sentiment Analysis", 2014 International Conference on Information Technology and Multimedia (ICIMU), November, 18 – 20, 2014, Putrajaya, Malaysia

[16] Vishal A. Kharde, S.S. Sonawane ,"Sentiment Analysis of Twitter Data: A Survey of Techniques", International Journal of Computer Applications (0975 – 8887) Volume 139 – No.11, April 2016

[17] Diksha Khurana1, Aditya Koli1, Kiran Khatter1,2 and Sukhdev Singh "Natural Language Processing: State of The Art, Current Trends and Challenges" ,Department of Computer Science and Engineering Manav Rachna International University, Faridabad-121004, India Accendere Knowledge Management Services Pvt. Ltd., India.

[18] Raja Selvarajan and Asif Ekbal, "IIT Patna: Supervised Approach for Sentiment Analysis in Twitter", Department of Computer Science and Engineering Indian Institute of Technology Patna, India.

[19] Prof. Alpa Reshamwala, Prajakta Pawar, Prof. Dhirendra Mishra, "REVIEW ON NATURAL LANGUAGE PROCESSING" IRACST – Engineering Science and Technology: An International Journal (ESTIJ), ISSN: 2250-3498, Vol.3, No.1, February 2013.

[20] Haque, Tanjim Ul, Nudrat Nawal Saber, and Faisal Muhammad Shah. "Sentiment analysis on large scale Amazon product reviews." 2018 IEEE international conference on innovative research and development (ICIRD). IEEE, 2018.

[21] Elmurngi, Elshrif Ibrahim, and Abdelouahed Gherbi. "Unfair reviews detection on amazon reviews using sentiment analysis with supervised learning techniques." J. Comput. Sci. 14.5 (2018): 714-726.
[22] Nandal, Neha, Rohit Tanwar, and Jyoti Pruthi. "Machine learning based aspect level sentiment analysis for Amazon products." Spatial Information Research 28.5 (2020): 601-607.
[23] Bhatt, A., Patel, A., Chheda, H., & Gawande, K. (2015). Amazon review classification and sentiment analysis. International Journal of Computer Science and Information Technologies, 6(6), 5107- 5110.
[24] Jagdale, Rajkumar S., Vishal S. Shirsat, and Sachin N. Deshmukh. "Sentiment analysis on product reviews using machine learning techniques." Cognitive Informatics and Soft Computing. Springer, Singapore, 2019. 639-647.
[25] Rain, Callen. "Sentiment analysis in amazon reviews using probabilistic machine learning." Swarthmore College (2013).