Data-Driven Decision-Making In E-Commerce: Leveraging Sentiment Analysis

Nishant Verma[1], Sumesh Sood[2]
[1, 2] Dept of Computer Science
[1, 2] Himachal Pradesh University, Shimla, Himachal Pradesh, India.

Abstract- *The rapid expansion of the e-commerce industry has inundated consumers with a plethora of choices, resulting in "e-confusion," encompassing similarity, overload, and clarity confusion. In developing countries like India, factors such as product price, ratings, reviews, and availability exacerbate this dilemma. This research project aims to develop a comprehensive application to assist users in making informed product selections, leveraging sentiment analysis to extract insights from customer reviews.Sentiment analysis, a vital natural language processing (NLP) method, plays a pivotal role across diverse domains by extracting valuable insights from user-generated data. It gauges attitudes, opinions, and emotional states conveyed by speakers or writers, enabling the assessment of vast datasets of social brand mentions.While sentiment analysis is a powerful tool, it is not without limitations, influenced by factors like sarcasm, irony, and cultural nuances. This paper addresses the research gap in applying sentiment analysis in e-commerce businesses, outlining a methodology encompassing web scraping, data cleaning, visualization, and sentiment analysis using VADER and ROBERTA models.The research collects and analyzes data from Amazon and Flipkart, emphasizing product variety, branding strategies, and price dynamics on these platforms. It underscores the importance of selecting the right sentiment analysis tool based on project goals.In conclusion, this research illuminates the intricate world of sentiment analysis in e-commerce, providing insights into product variety, branding strategies, price dynamics, and sentiment analysis nuances. The choice of sentiment analysis tool emerges as a pivotal factor shaping sentiment categorization, contributing to informed decision-making in the ever-evolving e-commerce landscape.*

Keywords- E-Commerce; Product reviews; RoBERTA; Sentiment analysis; VADER; Web-scrapping;

I. INTRODUCTION

In our increasingly digitized world, every action we take leaves a digital footprint, from online news consumption to mobile phone calls, social media interactions, and even grocery shopping with credit cards. This vast trove of data meticulously recorded and stored, serving diverse purposes from statistical analysis to personalized services [1]. Yet, the exponential growth of data and databases demands intelligent tools for transforming processed data into actionable information and knowledge. This paper delves into the pressing need for such tools, focusing on sentiment analysis as a critical component of data-driven decision-making.

Over the past decade, the rapid expansion of the e-commerce industry, culminating in a staggering 5.2 trillion USD in global e-retail sales in 2022, has presented consumers with an overwhelming abundance of choices [2]. This abundance has led to a phenomenon known as "e-confusion," encompassing three primary facets: similarity confusion, overload confusion, and clarity confusion. In developing countries like India, additional factors such as product price, ratings, reviews, and availability further compound this consumer dilemma. Recognizing the significance of informed consumer choices in fostering trust and satisfaction in online shopping, our research project addresses this challenge head-on. We aim to develop a comprehensive application to assist users in making well-informed product selections and identifying reputable sources for their purchases. Central to our approach is the incorporation of sentiment analysis, which provides an insightful assessment of customer sentiments and opinions expressed in product reviews.

Sentiment analysis is a vital natural language processing (NLP) method dedicated to the evaluation and categorization of sentiments or emotional expressions within textual content, including reviews, comments, social media posts, and more [3]. This technique plays a pivotal role across diverse domains such as marketing, customer service, and social media analysis. Opinion mining, as it is often termed, involves the extraction of valuable insights from user-generated data. Sentiment analysis draws upon a multifaceted toolkit encompassing computational linguistics, text analytics, and machine learning elements like latent semantic analysis, support vector machines, and Natural Language Processing.

The primary aim of sentiment analysis is to gauge the attitude, opinion, emotional state, or intended emotional communication conveyed by a speaker or writer. It serves as a valuable tool for assessing vast datasets of social brand mentions, enabling enterprises to filter content based on positive or negative comments, thereby pinpointing prevalent themes or issues that underlie the expressed sentiment.

Sentiment analysis, as a natural language processing technique, relies on computational algorithms to extract subjective information from text data, encompassing opinions, emotions, and attitudes [4]. Its core objective lies in discerning whether the expressed sentiment falls within the categories of positive, negative, or neutral, often assigning numerical scores to quantify the intensity of sentiment. Nevertheless, it is essential to acknowledge that sentiment analysis is not infallible and can be influenced by factors such as sarcasm, irony, and cultural subtleties.

II. BACKGROUND

The three underlying principles of e-confusion or customer confusion are similarity confusion, overload confusion and clarity confusion. But, in most developing countries like India, the biggest cause of consumer swingers' is product price, rating, reviews and of course availability. While sentiment analysis has been widely used in various industries, there is a need for more research on its application in e-commerce businesses to improve customer satisfaction and drive sales. Opinion mining was an agenda that exploited during this research project, it helped in determining whether the overall customer experience of the customer is positive, neural or negative in nature.

III. PROBLEM STATEMENT

A research gap in the lack of studies that focus on the use of sentiment analysis in e-commerce businesses specifically. Keeping this in mind, the project is aimed at building a thorough application to help users in deciding which product to buy from, and more importantly where from. In the context of an ongoing challenge, this project seeks to integrate data analytics methods, specifically incorporating KDD (Knowledge Discovery Database), with comparative analysis techniques to tackle the task of identifying the most suitable platform for purchasing specific products. Due to time limitations, the analysis will focus solely on Amazon and Flipkart. Subsequently, the project aims to collect and analyze product reviews, with a particular emphasis on Amazon, to perform sentiment analysis.

IV. TOOLS AND LIBRARIES

IDE Used: Python Shell, PyCharm, Jupyter Notebook
Python modules and libraries used: Beautiful Soup, requests, pandas, numpy, scipy, seaborn, random, matplotlib, warnings, time, datetime, regex, re,smtplib,Ipython,nltk.sentiment,tqdm.notebook,transformers, scipy.special (softmax), tabulate, urlib. request, PIL, Word count.

V. METHODOLOGY

The process and methodology involved during the project were a robust and in-depth exploration of various components and fragments of KDD and data analytics, primarily for the purpose of collecting and analyzing data from e-commerce websites (Amazon and Flipkart) and conducting sentiment analysis. While it may not follow a traditional research methodology typically found in academic research papers, it incorporates elements of data collection, data preprocessing, and data analysis commonly used in applied research and software development projects. Opinion mining was an agenda that was exploited during this research project, it helped in determining whether the overall customer experience of the customer is positive, neutral or negative in nature.

Here is a breakdown of the key components of the methodology:

A. **Designing the Architecture of the Program**: This likely involves defining the overall structure and components of the software program, which includes web scraping, data cleaning, visualization, and sentiment analysis.
B. **Web Scraping**: Two classes, one for Amazon and one for Flipkart, are implemented for web scraping. This involves collecting data from these e-commerce websites based on user-specified queries or products.
C. **Inspecting Elements and Trial and Error**: This step suggests that the methodology involves an iterative process of inspecting HTML elements on the Amazon and Flipkart websites and adjusting the web scraping process based on trial and error to extract the desired data.
D. **Cleaning and Transforming Data**: Data cleaning and transformation occur in real-time as the data scraped. This step likely involves removing noise and irrelevant information from the scraped data.
E. **Visualizing Product Range**: Visualizations created to compare the product range on Amazon and Flipkart for the specified product, providing insights into the availability and variety of products.
F. **Scraping Comments**: All comments or reviews related to the selected item scraped, likely for further analysis.
G. **Sentiment Analysis**: Two sentiment analysis models, VADER and ROBERTA, are used to analyze the sentiment of the scraped comments. This step involves

determining whether the comments are positive, negative, or neutral in sentiment.

H. **Visualizing Sentiment Analysis**: The results of sentiment analysis are visualized, possibly to provide a clear understanding of the sentiment distribution among the comments.

VI. DATA SOURCES AND COLLECTION

The impromptu response nature of the project required an on-demand and customizable dataset generation, every time the user seeks a product. To tackle the complexity, we created three modules of web scrapping for specific purposes and generating specific datasets.

The sources of the web-scrapped data were Amazon and Flipkart's website. The collection involved a generalized algorithm of web scrapping products and their information.

The reviews scrapped module went through all the comments of the product and, post data scrubbing, storing the information in a dataset. Post scrapping the reviews the sentiment analysis module created by me, added 8 columns based on a ROBERTA and VADER model sentiment scores. These scores helped in generating an aggregate opinion (polarity) representative of all the customers.

Two main datasets created through the process.

- Web Scrapping Amazon (based on query (product given by client)
- Web Scrapping Flipkart (based on query (product given by client)

The above datasets are merged and the final product dataset created with the web-scrapped data had the variables:

1. querry_searched
2. ecommerce_website
3. product_name
4. product_price
5. product_rating
6. rating_count
7. poduct_image_url
8. product_url

Web Scrapping Comments (based on the product selected from the above dataset) (Amazon Only).

Post applying NLP models like ROBERTA and VADER on the comments scrapped, to generate sentiment scores. The comments scrapped dataset had the variables:

1. customer_name
2. review_date
3. review_title
4. review_ratings
5. review_content
6. review_size
7. sentiment
8. roberta_sentiment
9. vader_sentiment
10. roberta_neg
11. roberta_neu
12. roberta_pos
13. vader_neg
14. vader_pos
15. vader_comp

VII. DATA ANALYSIS

Designing the architecture of the program

From the onset, it understood that the project required an elaborate backend and a sophisticated design to help integrate it with the finesse of the front end. This prompted me to create 3 separate modules and import it into the final file as .py finals. Using concepts from OOPs would result in a simplified and easy to trace back software.

The backend modules include:webscraping.py module:

-Web Scrapping Amazon Class (based on query (product given by the client)
-Web Scrapping Flipkart Class (based on query (product given by the client)

Dataset Created with the web scraped data had the variables:

1. querry_searched
2. ecommerce_website
3. product_name
4. product_price
5. product_rating
6. rating_count
7. poduct_image_url
8. product_url

-Web Scrapping Comments Class (based on the product selected from the above dataset)(Amazon Only)

Dataset Created with the web scraped data had the variables:

1. customer_name
2. review_date

3. review_title
4. review_ratings
5. review_content

ProjectVisualisations.py module:

- Consists of various functions that help in plotting the amazon vs Flipkart data.
- Using Plotly, seaborn, and matplotlib to come up with plots for a website vs website analysis.

ProductReviewAnalysis.py module:

This file contained was the main file for performing a sentiment analysis upon reviews/comments on a product (currently limited to amazon)

- Using ROBERTA Model (NLP)
- Using VADER Model (NLP)
- A weighted 60:40 approach (Roberta : Vader) to obtain a final sentiment value.

Inspecting elements of static HTML pages of Amazon and Flipkart and Scrapping based on trial and error.

The web scraping was done in tandem with the requests module and BeautifulSoup module, this involved requesting the search page based on the user's query. The page retrieved then had to be scouted for required information. This was done with basic knowledge of the inbuilt methods of the libraries being used and on a trial-and-error basis. It was a tedious process as it required checking where the elements were present on the actual page and comparing it in my script.

Cleaning data and transforming data while it's being scrapped

A major difference between traditional EDAs and models created by me in the past and this project was, I had to clean and make the data appropriate while it was being scrapped and before inserting it into the dataset. This involved using regex and string functions and logic to make the data presentable and comprehensible.

Visualizing Amazon vs Flipkart product range for the product

Using Plotly, seaborn, and matplotlib to paint an effective picture of how the products vary on Amazon and Flipkart for a user's query. I created various plots using user defined functions and recalled all of them under one visualizing parent function.

Scrapping All Comments based on the item selected

During the initial phase of scraping products from Amazon and Flipkart, the 'onclick()' URLs were also collected, anticipating their use in subsequent analysis. When selecting a product for sentiment analysis, the URL is employed to send a request. On the Amazon product details page, there is a 'see all reviews' button located at the bottom. A request is sent to the 'onclick()' URL of this button, and a loop is utilized to iterate through the page numbers, allowing for the extraction of all reviews, including the author's name, date, and star ratings. All of this information is systematically organized and stored in a pandas Data Frame

Sentiment Analysis using Vader and Roberta Model

The comments are analyzed with the help of NLTK module and hugging face models. I used two different models (VADER and Roberta) and a weighted 60:40 approach to extract an accurate sentiment for the product.

Visualizing Sentiment Analysis

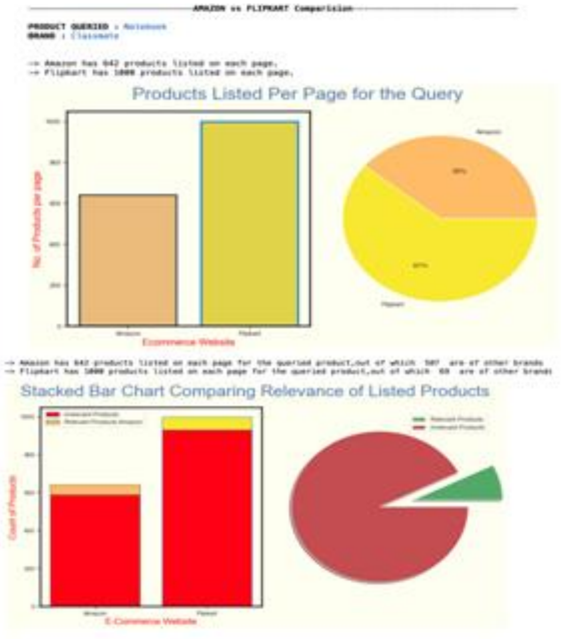
The following maps, graphs and comparisons are used to display the results of the sentiment analysis:

- Sentiment (Positive, Negative, Neutral) Count Plot, Pie Chart
- Weighted Sentiment Spread (distplot)
- Roberta vs Vader Model Sentiment Score comparison
- Review count for the product (Bar graph, Pie Chart)
- In depth comparison of Vader and Roberta Scores

VIII. OBSERVATIONS AND FINDINGS

AMAZON VS FLIPKART

The observations are on a per-case basis, as can be seen in the below Classmate Notebook Comparison the results are self-explanatory.



```
------------------------AMAZON vs FLIPKART Comparision------------------------

PRODUCT QUERIED : Notebook
BRAND : Classmate

-> Amazon has 642 products listed on each page.
-> Flipkart has 1000 products listed on each page.
```

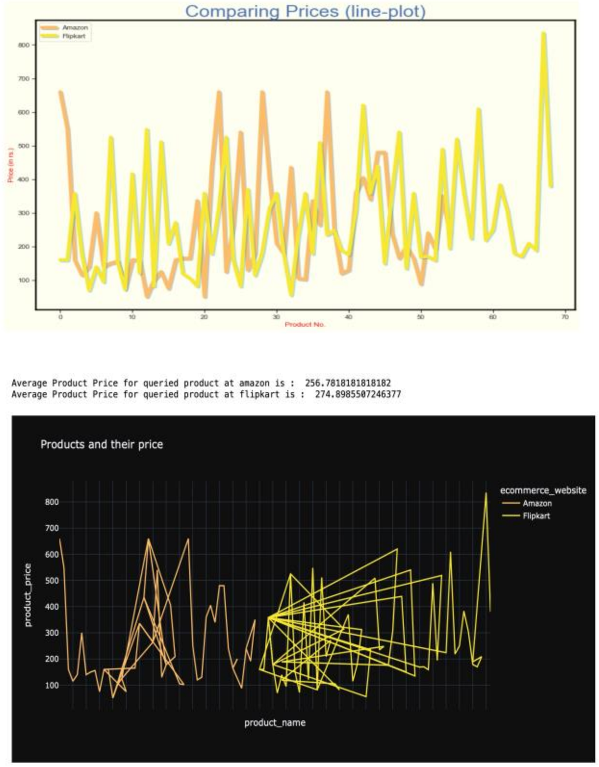
```
-> Amazon has 642 products listed on each page for the queried product,out of which  587  are of other brands
-> Flipkart has 1000 products listed on each page for the queried product,out of which  69  are of other brands
```

Figure 1: Products per-page comparison of amazon vs Flipkart

In the above comparison (Figure: 1), when comparing the two platforms, it becomes evident that Amazon features 642 products per page, with 587 of these products coming from brands other than Amazon itself. In contrast, Flipkart displays 1000 products on each page, with a much smaller subset of 69 products attributed to brands other than Flipkart.

This comparison highlights that Amazon's product listings primarily consist of products from external brands, while Flipkart tends to highlight a larger proportion of its own branded products on its pages. This distinction suggests differing approaches to product selection and branding between these two e-commerce platforms.

The average product price for the queried product on Amazon is 256.78, whereas the average product price for the same queried product on Flipkart is slightly higher at 274.89 (Figure: II). This discrepancy in average prices attributed to several factors. The difference in average product prices between Amazon and Flipkart can be influenced by various factors related to product selection, pricing strategies, and seller dynamics. Profitability for each platform is determined by a combination of factors, including profit margins, sales volume, additional revenue streams, and operational costs. The specific financial performance of each platform would require a more detailed analysis of their financial statements and business strategies.



```
Average Product Price for queried product at amazon is :  256.7818181818182
Average Product Price for queried product at flipkart is :  274.8985507246377
```

Figure II: Price comparison of Amazon vs Flipkart

SENTIMENT ANALYSIS USING VADER AND ROBERTA MODEL

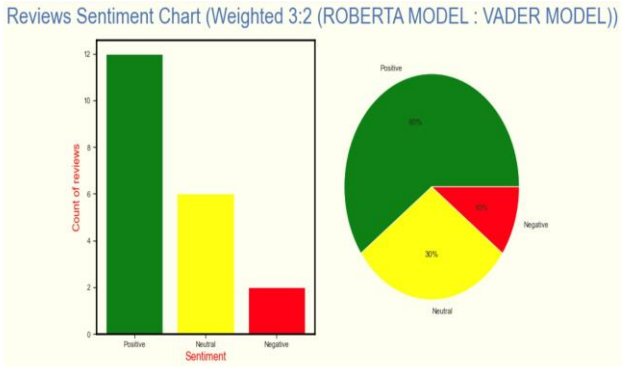


Figure III: Total count of product reviews

By looking at the heights of these bars (Figure: III), you can quickly assess the distribution of sentiment in the dataset. Here, the "Positive" bar is the tallest, it means that most reviews in the dataset are positive in sentiment and the "Negative" bar is the smallest, it suggests that the majority of reviews are positive. The height of the "Neutral" bar represents the count of reviews that are neither strongly positive nor negative. These reviews typically fall in the middle, indicating a lack of strongly polarized sentiment.

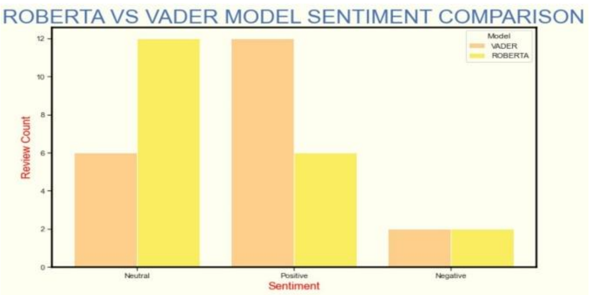


Figure IV: Roberta vs Vader model sentiment comparison

In essence, this above (Figure: IV) data demonstrates how RoBERTa and VADER categorize sentiments differently for the same dataset. RoBERTa tends to identify more instances as positive and fewer as neutral, whereas VADER classifies more instances as neutral and fewer as positive. Both tools agree on the classification of negative sentiments.

These results underscore that the choice of sentiment analysis tool can significantly influence the distribution of sentiment categories, and the selection of the tool should align with the specific requirements and goals of the sentiment analysis task.

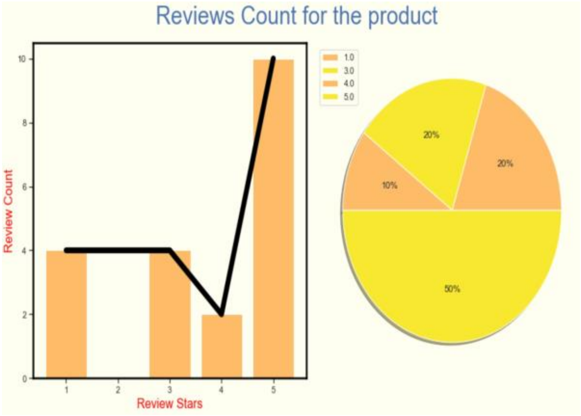


Figure V: Product rating vs Review length

The graph provides a comprehensive overview of review counts for a specific product, categorized by 5-star ratings (Figure: V). Each bar on the graph corresponds to a particular rating, ranging from one to five stars; with the height of each bar representing the total number of reviews awarded that specific rating. This visual representation allows us to gain valuable insights into customer sentiment and feedback.

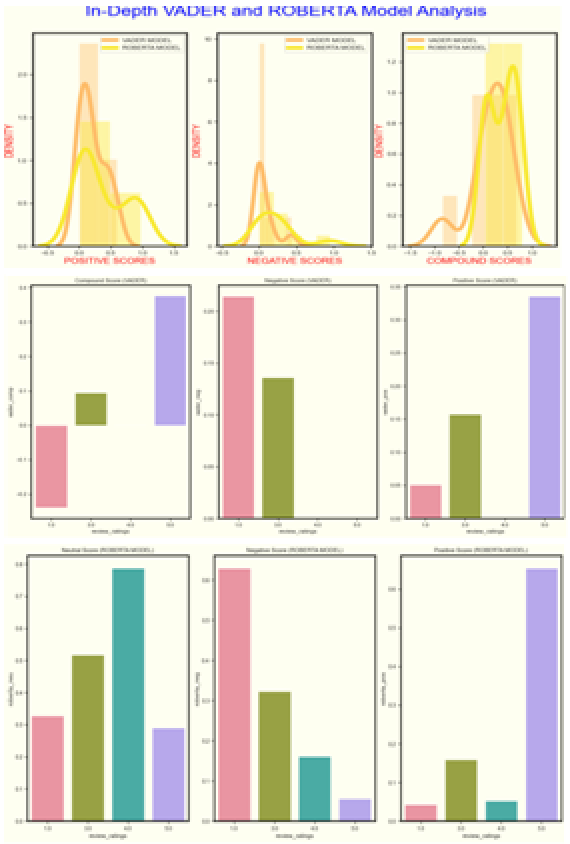


Figure VI: In-depth Vader and Roberta model analysis

IX. CONCLUSION

The culmination of this research journey has brought to light a multitude of insights into the intricate world of sentiment analysis in the realm of e-commerce. This chapter succinctly encapsulates the primary outcomes of this study:

- **Product Variety and Branding Strategies:** An exploration of two e-commerce titans, Amazon and Flipkart, has uncovered profound distinctions in their product selection and branding strategies. Amazon, with its expansive range of products primarily from external brands, juxtaposed Flipkart's more focused approach, with a substantial portion attributed to in-house brands.
- **Price Dynamics:** A comprehensive analysis of product prices across Amazon and Flipkart unveiled nuanced disparities. While Amazon presented a lower average product price for the queried item, these variations stem from intricate factors, including product assortment, competitive pricing strategies, seller diversity, and promotional activities. It is imperative to recognize that profitability extends beyond product prices, encompassing diverse facets like profit margins, sales volume, and operational costs.

- **Sentiment Analysis:** Employing sentiment analysis techniques, including RoBERTa and VADER models, to scrutinize customer reviews has unearthed a panorama of customer sentiments. RoBERTa and VADER exhibited disparities in sentiment categorization, with RoBERTa predominantly identifying more positive sentiments and VADER emphasizing a higher proportion as neutral. The selection of the sentiment analysis tool emerged as a pivotal factor shaping the distribution of sentiment categories.

X. FUTURE SCOPE

The culmination of this research chapter opens avenues for future investigations, including expanding the analysis to encompass a wider array of e-commerce platforms, seeking proprietary data access for deeper insights, exploring diverse sentiment analysis tools and models, conducting longitudinal analyses to track evolving customer sentiments, and integrating multiple data modalities for a comprehensive view of customer feedback. In conclusion, this research illuminates the dynamic landscape of e-commerce, unveiling customer sentiments and offering a strategic roadmap for success in the digital commerce sphere, where the ability to decipher and respond to customer sentiments remains paramount for continued growth and prosperity.

REFERENCES

[1] R. Ahuja, A. Chug, S. Kohli, S. Gupta, and P. Ahuja, "The impact of features extraction on the sentiment analysis," Procedia Computer Science, vol. 152, pp. 341-348, 2019.

[2] S. Aravindan and A. Ekbal, "Feature extraction and opinion mining in online product reviews," in 2014 International Conference on Information Technology, pp. 94-99, IEEE, December 2014.

[3] T. Arora, R. Gupta, and P. Goyal, "Sentiment analysis using machine-learning techniques: A review," Journal of Computer Science and Engineering, vol. 1, no. 1, pp. 7-17, 2018.

[4] Y. Chandra and A. Jana, "Sentiment analysis using machine learning and deep learning," in 2020 7th international conference on computing for sustainable global development (INDIACom), pp. 1-4, IEEE, March 2020.

[5] X. Fang and J. Zhan, "Sentiment analysis using product review data," Journal of Big Data, 2015.

[6] S. S. Hanswal, A. Pareek, G. Vyas, and A. Sharma, "Sentiment analysis on E-learning using machine learning classifiers in Python," in Rising Threats in Expert Applications and Solutions: Proceedings of FICR-TEAS 2020, Springer Singapore, 2021.

[7] R. S. Jagdale, V. S. Shirsat, and S. N. Deshmukh, "Sentiment Analysis on Product Reviews Using Machine Learning Techniques," in Cognitive Informatics and Soft Computing, Advances in Intelligent Systems and Computing, Springer, Singapore, 2019.

[8] Z. Kastrati, F. Dalipi, A. S. Imran, K. PirevaNucprovide, and M. A. Wani, "Sentiment analysis of students' feedback with NLP and deep learning: A systematic mapping study," Applied Sciences, 2021.

[9] W. Liao, B. Zeng, X. Yin, and P. Wei, "An improved aspect-category sentiment analysis model for text sentiment analysis based on RoBERTa," Applied Intelligence, 2021.

[10] L. Mathew and V. R. Bindu, "A review of natural language processing techniques for sentiment analysis using pre-trained models," in 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), IEEE, March 2020.

[11] K. L. Tan, C. P. Lee, K. S. M. Anbananthen, and K. M. Lim, "RoBERTa-LSTM: a hybrid model for sentiment analysis with transformer and recurrent neural network," IEEE Access, 2022.

[12] C. Slamet, R. Andrian, D. S. A. Maylawati, D. Suhendar, W. Darmalaksana, and M. A. Ramdhani, "Web scraping and Naïve Bayes classification for job search engine," in IOP Conference Series: Materials Science and Engineering, IOP Publishing, January 2018.

[13] P. V. Rajeev and V. S. Rekha, "Recommending products to customers using opinion mining of online product reviews and features," in 2015 International Conference on Circuits, Power and Computing Technologies [ICCPCT-2015], IEEE, March 2015.