Indian Journal of Information Sources and Services
ISSN: 2231-6094 (P) Vol.14, No.3, 2024, pp.52-63
© The Research Publication, www.trp.org.in
DOI: https://doi.org/10.51983/ijiss-2024.14.3.08

Harnessing the Power of Web Scraping and Machine Learning to Uncover Customer Empathy from Online Reviews

A. Anny Leema[1], Dr.P. Balakrishnan[2*] and N. Jothiaruna[3]

[1]Associate Professor Sr., Analytics Department, School of Computer Science and Engineering, VIT Vellore, India
[2*]Associate Professor Sr., Analytics Department, School of Computer Science and Engineering, VIT Vellore, India
[3]Assistant Professor, Data Science Department, School of Computing Science Engineering and Artificial Intelligence, VIT Bhopal, India
E-mail: [1]annyleema.a@vit.ac.in, [2]balakrishnan.p@vit.ac.in, [3]jothiaruna@vitbhopal.ac.in
ORCID: [1]https://orcid.org/0000-0002-0704-2794, [2]https://orcid.org/0000-0002-2960-636X, [3]https://orcid.org/0000-0001-5098-1845

(Received 02 May 2024; Revised 14 June 2024, Accepted 18 July 2024; Available online 30 September 2024)

Abstract - **In today's information-driven world, companies need to grasp customer sentiment and pinpoint key product features to make informed choices and improve customer satisfaction. The abundance of online customer feedback provides businesses with a rich source of information about customer perceptions and preferences. Yet manual analysis of large volumes of reviews is a tedious and resource-intensive task. It describes a web scraping application that sorts reviews into positive, negative, or neutral sentiments that businesses can use for insight. By assigning weights to reviews based on perceived impact, the approach offers a nuanced understanding of customer opinions. Experimental results demonstrate its effectiveness in accurately categorizing and analysing reviews, distinguishing between genuine and fake feedback. Additionally, AI-powered examination of product reviews employs NLP methods to extract important information from customer feedback, identifying particular elements that customers like or find troubling. After the successful identification of gain points, they can be used to improvise product development and marketing. So, product-specific focused crawler is developed to extract customer reviews from the most trusted websites. As a side chain, this can also be used as a tool for analysing the competitors' products to understand their pain as well as gain points. The approach can uncover the most discussed aspects of products, enabling businesses to better understand customer perceptions and preferences to make better decisions thereby saving time and money.**

Keywords: **Web Scraping, Sentiment Analysis, Natural Language Processing, Data Analysis Algorithms, AI Trends, Marketing Strategies**

I. Introduction

The recent decade has witnessed a surge in e-commerce companies, including Amazon, Flipkart, Myntra, Shopify, and eBay. The global e-commerce industry is currently valued at $3.1 trillion and is projected to reach $50 trillion by 2050. The COVID-19 pandemic accelerated the adoption of online shopping as consumers sought the convenience and cost-effectiveness of purchasing goods from the comfort of their homes (Sunarto et al., 2023). With a few clicks, they can order a wide range of products, from clothing and groceries to appliances and electronics. However, not all products available on these platforms are authentic, leading many customers to fall prey to dishonest sellers making false claims about their products' authenticity. Consequently, online reviews have become indispensable for consumers seeking trustworthy information before making purchasing decisions (Abdo et al., 2023). Businesses, in turn, have recognized the importance of these reviews in understanding market perception of their products. Despite their significance, manually analysing a large volume of reviews is a labor-intensive and time-consuming endeavor.

Making use of customer feedback and product specifications provides customers with valuable insights into the quality and perceived worth of a product, aiding them in making informed decisions when making a purchase. We suggest a Python application that analyses product reviews to offer a summary of the product using subjective criteria. The complete review could potentially influence customers' decisions on whether to buy the product or not. This encompasses a variety of items such as electronics, apparel, home goods, groceries, and vehicles purchased through online channels. This is necessary to ensure the authenticity of the product and also enables customers to make educated decisions. Automated techniques have been created to extract sentiment details from online reviews. Product reviews on popular websites are categorized based on sentiment using NLP techniques. By revealing customer attitudes, such data are able to aid companies in the creation of new products or in marketing (Liu, 2012).

The suggested approach utilizes sophisticated algorithms to categorize reviews as either positive, negative, or neutral in order to provide businesses with insight into customer opinions. It also considers the impact of each review in order to provide a more detailed view of customer opinions. We make use of sophisticated NLP methods to analyze customer reviews and also offer businesses valuable insights into customer sentiment. NLP also recognizes the positive and negative aspects of specific product web crawlers, known as "gain points" and "pain points" respectively. We make use of this data to enhance our products and to gain a competitive advantage by analysing competitor products. Businesses can increase efficiency and reduce costs by understanding their customers' preferences and dislikes. Briefly, AI-driven review analysis helps businesses understand customer preferences better, leading to more informed decision-making. Research has demonstrated that our method effectively categorizes reviews with precision and offers practical information (Oleksandr et al., 2024). Our method can assist companies in various sectors in comprehending customer attitudes, enabling them to make informed choices regarding product enhancement and marketing strategies based on data.

II. Background Study and Related Work

Today, marketing, product development, and customer care are overwhelmed with unstructured data from customer reviews. Businesses wanting to comprehend customer opinions encounter difficulty in extracting valuable insights from these assessments. Existing methods usually involve time-consuming manual processes or basic algorithms that lack the ability to comprehend the meaning behind assessments. As a result, companies miss out on chances to identify the strengths and weaknesses of their products, which could hinder their ability to enhance services and meet customer needs (Yemunarane et al., 2024). There is a pressing need for the creation of a smart AI system which can gather and assess user feedback in order to offer valuable suggestions for improving products.

The integration of artificial intelligence in business has revolutionized the process of analysing product reviews, leading to a surge in research focused on extracting valuable customer insights. Scholars highlight the importance of this analysis, pointing to the work of (Htay & Lynn, 2013), who utilize linguistic rules and part-of-speech tagging to extract valuable insights from reviews. In the same way, C. Chauhan and colleagues (Chauhan & Sehgal, 2017). Analysing stress sentiments on platforms such as Flipkart and Amazon is crucial in order to improve products and marketing tactics by using web scraping and polarity analysis. Zhang et al., (2022) created a system for choosing products using online reviews that combines sentiment analysis and also the intuitionistic fuzzy - TODIM technique (Zhang et al., 2022). Researchers create advanced techniques such as the SLCABG model for sentiment analysis. Yang et al., utilized sentiment lexicons and deep learning models to analyze complex emotions expressed in reviews (Yang et al., 2020). Deep learning techniques, exemplified by artificial intelligence. Parvez and colleagues, as well as Zhang, emphasize that these developments indicate a growing sentiment analysis field that is essential for obtaining business insights through AI (Parvez et al., 2018; Abdo et al., 2023; Zhang et al., 2018).

Sentiment analysis isn't limited to product reviews. Making use of Bi - directional LSTM in combination with CNN models to identify entities in e-commerce (Shah et al., 2022). In 2008, A.N. Muhammad and his associates conducted sentiment analysis on comments posted on YouTube. Its ability to adapt and influence the comprehension of consumer opinions across various platforms is demonstrated in (Muhammad et al., 2019). This work focuses on the significance of sentiment analysis in utilizing AI for more comprehensive business understanding. Strategies for customer engagement. Provides an overview of previous research in the field of artificial intelligence as it pertains to analysing customer reviews, sentiment analysis, along with various applications shown in table I.

TABLE I PROVIDES AN OVERVIEW OF PREVIOUS RESEARCH IN THE FIELD OF ARTIFICIAL INTELLIGENCE AS IT PERTAINS TO ANALYSING CUSTOMER REVIEWS, SENTIMENT ANALYSIS, ALONG WITH VARIOUS APPLICATIONS

Topic	Key Contributions	Authors	Year	Reference
Linguistic rules and POS tagging for insights	Leveraged linguistic rules and part-of-speech tagging to extract meaningful insights from customer reviews.	(Htay & Fong, 2013)	2013	[28]
Sentiment analysis on e-commerce platforms	Emphasized the importance of sentiment analysis on platforms like Amazon and Flipkart to improve products and marketing strategies.	(Chauhan et al., 2019)	2019	[11]
Sentiment analysis with fuzzy TODIM method	Proposed a method for product selection based on reviews using sentiment analysis and intuition.	(Zhang et al., 2020)	2020	[29]
SLCABG model for sentiment analysis	Developed the SLCABG model integrating sentiment lexicons and deep learning architectures to capture nuanced sentiments.	(Yang et al., 2020)	2020	[13]
Deep learning in sentiment analysis	Investigated the usage of rich learning architectures for enhancing sentiment analysis accuracy.	(Parvez et al., 2021; (Zhang et al., 2021)	2021	[14], [15]
Entity detection in e-commerce	Applied bidirectional LSTM with CNN models for entity detection in e-commerce contexts.	(Shah et al., 2021)	2021	[16]
Sentiment analysis on YouTube comments	Showed how sentiment analysis can be used on comments from YouTube to fully grasp consumer sentiment across various platforms.	(Muhammad et al., 2022)	2022	[17]

The World Wide Web (WWW) is a massive information resource that houses zetabytes of data and is used by an increasing number of internet users who are looking for a variety of online resources. In this situation, having efficient search engines is crucial for quickly and efficiently accessing data while using minimal resources. These search engines enable users to efficiently browse a vast and varied array of information by organizing pages based on subject and importance. In essence, a search engine consists of three main components: web crawlers that locate and categorize webpages, a database that stores indexed information, and search interfaces that process user queries to help locate specific content efficiently.

A. *Web Crawling*

Web crawlers assist search engines in locating information by scanning data across numerous websites that are spread out. Spidering, also known as web-spying, is a computerized procedure that scans the internet for data and saves it for future access. Search engines are built on a technique that enables them to organize and prioritize web pages, helping users find their way through the vast amount of information available online.

The seeds consist of a collection of established websites that serve as the starting point for web crawlers. They proceed to click on the links provided on those sites to be able to discover additional websites. They then add the new website to their crawl frontier list of sites to visit once they have found it. The main loop of the Crawler consistently chooses the next URL on the Frontier, downloads the associated webpage using HTTP, scans the downloaded page for embedded URLs, includes these newly found URLs in the Frontier, and saves the initial page in a nearby repository. The web crawlers store data from each site they access when they are archiving sites. This data can be utilized in order to generate a copy of the website as a precautionary measure, or to monitor the evolution of the website over a period of time. The internet relies on web crawlers for much of its functionality. They simplify the process of maintaining a well-organized and easily navigable online environment for individuals. Searching for info on the web would be impossible without the use of a web crawler. In order to find the information needed, Web Crawler utilizes the breadth first search algorithm or the depth first search algorithm.

B. *Classification of Web Crawlers*

Web crawlers play a crucial role in various applications, including search engines, price comparison websites, and research initiatives. As the core component of a vertical search engine, developing more accurate and efficient crawlers for information retrieval has become a crucial research area in the field of crawlers (Hati et al., 2010). This topic has captivated the attention of numerous researchers worldwide.

Different types of web crawlers can be classified based on their crawling strategies, target content, and underlying techniques.

1. *General-purpose Crawlers*

These web crawlers search and gather information from a variety of sites and domains across the internet. Search engines like Bing and Google usually utilize them to create extensive databases of websites.

2. *Focused Web Crawler*

It locates and gathers online pages that are connected to a specific subject, field, or category of information. Focused crawlers avoid irrelevant pages and focus on specific pages instead of using general purpose crawlers to index the whole internet. This method is especially effective for specialized search engines, comparison websites for prices, and other projects that involve in-depth exploration of a particular subject. Specialized web crawlers have a number of benefits, such as reduced hardware and network resource expenses, limited network congestion, and comprehensive search capabilities (Pan et al., 2019; Pavalam et al., 2012; Mali & Meshram, 2012; Kausar et al., 2013; Dey et al., 2010).

3. *Incremental Crawler*

These web crawlers prioritize updating current search results by recognizing any new or changed content that has been added after the previous crawl. They are useful especially when you are taking a look at dynamic websites that are likely to be updated with new content frequently.

4. *Personalised Crawler*

These are crawlers built for particular users or organisations. They can collect specific information, track personal interest or monitor competitor activities.

5. *Deep Crawler*

It is able to look at the inner workings of a website and uncover its various subpages and links. They are primarily utilized for gathering competitive intelligence, discovering content, and conducting in-depth website analysis.

6. *Distributed Crawler*

The internet is so large that one crawler process can not possibly retrieve all its data. For large websites or website networks a distributed crawler is an option. A distributed crawler uses several computers to share the crawling load. This enables faster crawling &amp; crawling of larger sites. They are typically more expensive than parallel crawlers but more scalable. Currently, the enhanced UbiCrawler is an example of a distributed crawler which can work on different network types and significantly shorten the crawling time (Kamoonpuri & Sengar, 2023).

7. *Parallel Crawler*

Parallel web crawlers are programs that make use of multiple threads or processors to be able to speed up the crawling process. Using multiprocessing or multithreading enables the crawler to distribute tasks among multiple separate processes or threads, functioning as individual crawlers. These crawlers share two major data structures to facilitate collaboration: The borderland and the storage facility. The frontier contains a collection of URLs waiting to be visited. The repository contains URLs which have been indexed. A designated frontier manager synchronizes the shared data structures to prevent conflicts when multiple writes are taking place simultaneously. Using a synchronized approach can greatly speed up the crawling process of parallel web crawlers, leading to faster internet traversal times.

III. Web Scraping

Web scraping involves automatically extracting information from websites. So it involves writing a program to send a request to a website, and then parsing the HTML of that site to get the information retrieved. This information is then used for data analysis, research or content aggregation. Figure 1: Web scraping - extracting data from web sites into structured data. This could be for market research / price comparison / lead generation / content aggregation etc.

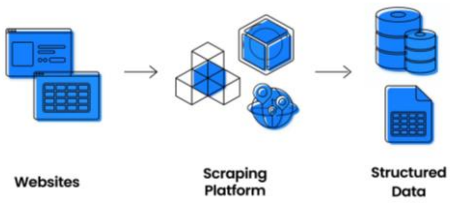


Fig. 1 Web Scraping Analyzes and Processes Online Information and Turns it into Structured Data

Web scraping is possible manually by copying and pasting data from websites but this is a slow and error-prone process. Robotic web scraping using special software or programming languages like JavaScript is far more effective and accurate. For the automation we have JavaScript library called "cheerio.js," a popular Node.js library for web scraping. It's a light and fast library for web scraping with a simple API for querying and manipulating HTML/XML documents.

TypeScript is a Microsoft open-source programming language. It is a typed superset of JavaScript that contains optional static typing, classes and interfaces. TypeScript can be used to write big web applications. TypeScript adds static typing to JavaScript for developers - improving code readability, tooling support, and code maintainability. TypeScript also lets developers spot possible errors and bugs at compile time instead of runtime, saving time and improving code quality.

TypeScript is compatible with JavaScript code. TypeScript code is compiled into JavaScript. TypeScript also supports modern JavaScript features like async/await and arrow functions for web development.

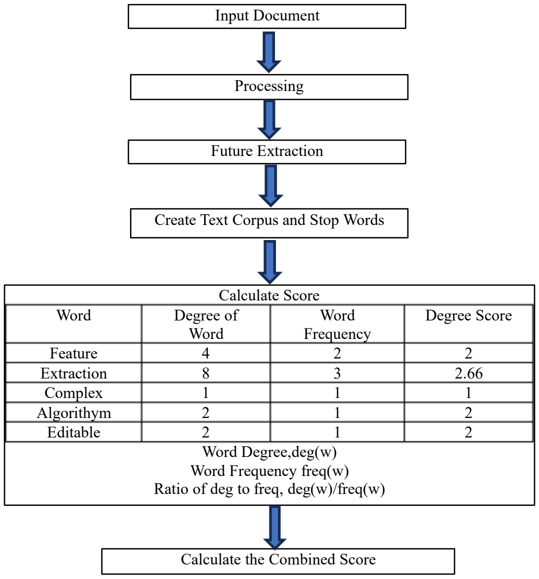
A. *Sentimental Analysis*

It is also called opinion mining. The sentiment analysis is used for market research, customer feedback analysis and social media monitoring. Sentiment analysis was performed in our web scraping application with a Natural Language Processing (NLP) model provided by the Spacy library in Python. Spacy provides tools and models for named entity recognition, part-of-speech tagging and sentiment analysis in an open-source library for NLP.

Using the pre-trained NLP model in the Spacy library, we quickly and accurately analyzed the sentiment of the customer reviews scraped from Amazon. This model combines machine learning with rule-based systems to identify sentiment in text data.

B. *Keywords Extractions*

Keyword extraction allows us to identify the most appropriate words within the text that express the author's emotions. We utilized the RAKE algorithm to extract keywords from the text. The text is segmented into separate words and then organized based on their importance to the overall message of the text. This is achieved by calculating a rating for every word using two criteria. Next, we utilized the spaCy library in Python to extract a collection of important tokens from each abstract by analyzing their frequency within the text and their relationships with other key words. This proved to be highly beneficial for the final stage of the topic modeling process. A collection of numerous tokens can be created by removing tokens that occur less frequently. This collection of texts was used as a reference when generating fresh topics. The total score of potential key phrases identified by the RAKE algorithm is determined by a combination of their frequency and significance, as demonstrated in Figure 2. Frequency refers to the number of times a specific phrase is repeated within the text. Higher scores are given to phrases that are frequency dependent.


Input Document
Processing
Future Extraction
Create Text Corpus and Stop Words
Calculate Score

Word	Degree of Word	Word Frequency	Degree Score
Feature	4	2	2
Extraction	8	3	2.66
Complex	1	1	1
Algorithym	2	1	2
Editable	2	1	2

Word Degree,deg(w)
Word Frequency freq(w)
Ratio of deg to freq, deg(w)/freq(w)
Calculate the Combined Score

Fig. 2 Rapid Automatic Keyword Extraction (RAKE) Algorithm for Keywords Extraction from Product Reviews

The number of consecutive words in a phrase is its degree. A phrase of higher degree is considered more significant and receives a higher score. A phrase's final cumulative score is its frequency multiplied by its degree. High cumulative scores mean more important phrases. The formula for computing the cumulative score of a phrase is:

$$\text{Frequency} * \text{Degree} = \text{Cumulative Score}$$

This cumulative score is used by RAKE to rank candidate key phrases and select the relevant ones as keywords. It has been shown that RAKE can identify relevant keywords in a wide range of contexts, including social media posts and academic research papers. Using the RAKE algorithm for each product review, we identified relevant keywords to understand customer sentiment. For RAKE implementation, we used Python programming language and the natural language toolkit (NLTK) library containing various tools and techniques for natural language processing. And we used Pandas library to process Amazon - sized data sets.

The candidate keywords are then sorted by score in RAKE, and the top N keywords are the output. This N value can be varied depending on the application. Here we applied the RAKE algorithm to scrape keywords from Amazon product descriptions and reviews. This enabled finding relevant keywords for each product which we could use for further analysis or visualization.

C. Python Microservice Development

Microservices architecture breaks a complex application into smaller, standalone services that perform different functions or responsibilities. It is a method different from conventional monolithic architectures where the whole application is built as one unit. Microservices enable easier development, deployment, and maintenance by isolating components, which improves scalability, flexibility, and fault tolerance (Boldi et al., 2017). Microservices architecture aims at enabling fast and scalable sentiment analysis and keyword extraction tasks. Sentiment analysis involves evaluating text to determine whether the sentiment is positive, negative, or neutral, and keyword extraction involves finding significant terms or phrases that represent topics in the text data. Python is selected for its flexibility and rich library support for information processing and machine learning tasks. FASTAPI is a contemporary web framework for developing Python APIs able to perform well with asynchronous tasks. python and Fastapi in a Microservices architecture create a scalable, flexible and efficient system for sentiment analysis and keyword extraction tasks.

IV. AI-POWERED ANALYSIS FOR IDENTIFYING PAIN POINTS AND GAIN POINTS

The system being suggested for real-time evaluation of customer sentiments in product reviews on an E-commerce platform is split into two main phases. The initial phase involves creating a sentiment analysis model, while the second phase involves making use of this model to analyze product reviews in real-time. It is important to consider pagination because many websites divide reviews into multiple pages. It is crucial to identify pagination links or buttons and incorporate a system into the scraping procedure to navigate through these review pages. In order to stop the website server from becoming overwhelmed and reduce the chances to be blocked, it is crucial to put into action rate limiting and practice respectful behavior within the web scraper. This involves controlling how often and how much requests are made. The extracted information can be saved for later examination in JSON or CSV file formats, or it can be directly uploaded into a database. The Seq2Seq model is especially beneficial for summarizing sentences because it is able to process input sequences of different lengths and generate summaries of different lengths. Sentiment extraction is carried out using LSTM model. Figure 3-a illustrates the sq2seq architecture for Text Summarization, while Figure 3 - b shows the exact same architecture. This visualization and breakdown of the algorithm show how the Embedding Layer and LSTM Layers collaborate to analyze sentiment in text inputs, transforming them into accurate sentiment predictions. The three main tasks demonstrated are:

The Embedding Layer converts words into uninterrupted vectors.

LSTM Layers analyze embedded sequences in order to forecast sentiment labels.

Predicted Sentiment Label: The result from the LSTM model that indicates the predicted sentiment.

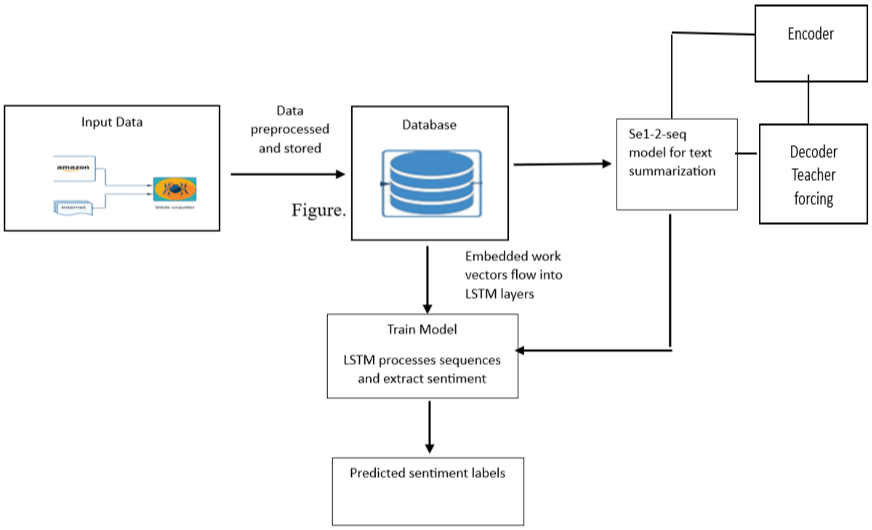


Fig. 3-a Text Summarization and Sentiment Label Prediction

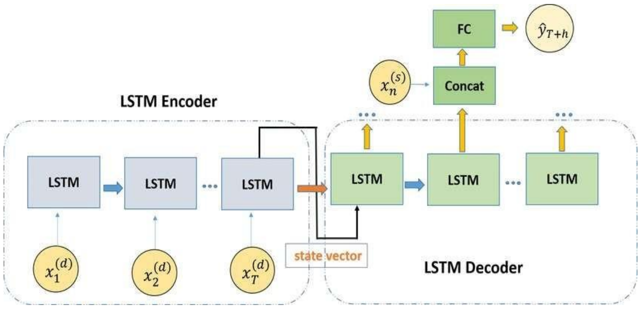


Figure 3-b Seq2Seq Architecture

The proposed system uses a Seq2Seq model with an encoder-decoder architecture for text summarization. The encoder captures the context of the input text (review) and the decoder generates a summary based on that context. Teacher forcing is used during training to improve the model's performance. It helps to stabilize and accelerate the training of Seq2Seq models by providing consistent and correct inputs during the learning process.

A. *Sentence Sentiment Extraction*

Assessing product qualities in online e-commerce is complex due to the varied expressions used by reviewers. Identifying nuanced sentiments, especially those expressed through negative prefixes, is challenging yet crucial for accurate analysis. For instance, phrases like "not wrong" may convey less negative sentiments despite containing negative terms. To address this, a Negative Polarity Identification algorithm is employed to pinpoint such expressions, enhancing sentiment analysis accuracy ChatGPT

In this method, the Seq2Seq model produces summaries, which are subsequently fed into an LSTM model to extract sentiments from reviews. The LSTM employs an embedding layer to transform word representations into continuous vectors, learning associations from the data it's trained on. Using a sequential architecture, LSTM networks include gates (input, output, forget) to control the flow of information within and between memory cells. At each step in the sequence, these gates handle input, retention of past information, and output, enhancing the analysis of sentiments and evaluation of products in e-commerce contexts.

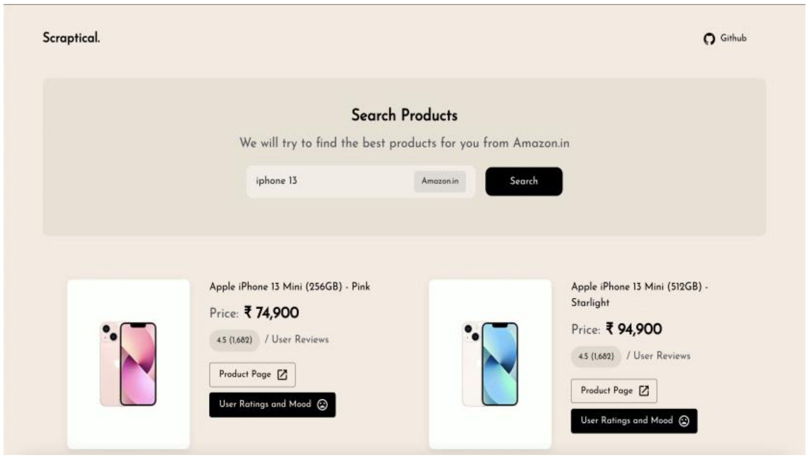
$$FT=\sigma(W_f*[h_{t-1},x_t]+B_f) \qquad \text{Equation (1)}$$

In this equation (1), the forget gate's value FT is influenced by both the current input x_t and the previous output h_{t-1} Its value ranges between 0 and 1, where 0 signifies that the LSTM should disregard the previous information, while 1 indicates that it should retain it. During training, the model optimizes its parameters using an appropriate loss function. For multi-class sentiment analysis, categorical cross-entropy is typically used, while binary cross-entropy is common for binary sentiment analysis. The training dataset consists of reviews paired with their corresponding sentiment labels.

Understanding product reviews in online stores can be tricky. Reviewers use a wide variety of words and phrases, and negative prefixes like "not" can make it difficult to tell if a review is positive or negative. For example, "not bad" is less negative than "bad." To address this challenge, we built a system that can identify these nuanced phrases and analyze the sentiment of reviews. The proposed system identifies Negation, Summarizes Reviews and analyse sentiments.

B. *Client Side Application*

The user interface (UI) prioritizes a smooth user experience. Material-UI (MUI) was chosen for pre-built and customizable components to create a modern and responsive design. TypeScript, a superset of Javascript, was used for catching errors during development and improved maintainability. Next.js, a React-based framework, offered features like server-side rendering for a fast and SEO-friendly UI as shown in Figure 4. Finally, loading skeletons were implemented to show users data is being retrieved while keeping the UI responsive. Overall, the UI development prioritized ease of use, popularity within the React community, and resulted in a modern and responsive interface. The sample screenshot for product reviews is depicted in Figure 5.



Search Products

We will try to find the best products for you from Amazon.in

iphone 13

Apple iPhone 13 Mini (256GB) - Pink

Price: ₹ 74,900

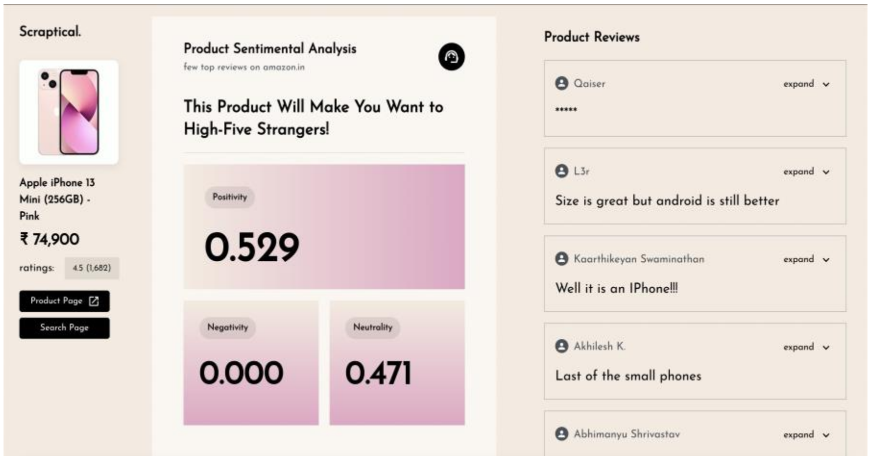
4.5 (1,682) / User Reviews

Apple iPhone 13 Mini (512GB) - Starlight

Price: ₹ 94,900

4.5 (1,682) / User Reviews

Fig. 4 Search Page Showing Results for Iphone 13



Scraptical.

Apple iPhone 13 Mini (256GB) - Pink

₹ 74,900

ratings: 4.5 (1,682)

Product Sentimental Analysis

few top reviews on amazon.in

This Product Will Make You Want to High-Five Strangers!

Positivity 0.529

Negativity 0.000

Neutrality 0.471

Product Reviews

Qaiser
•••••

L3r
Size is great but android is still better

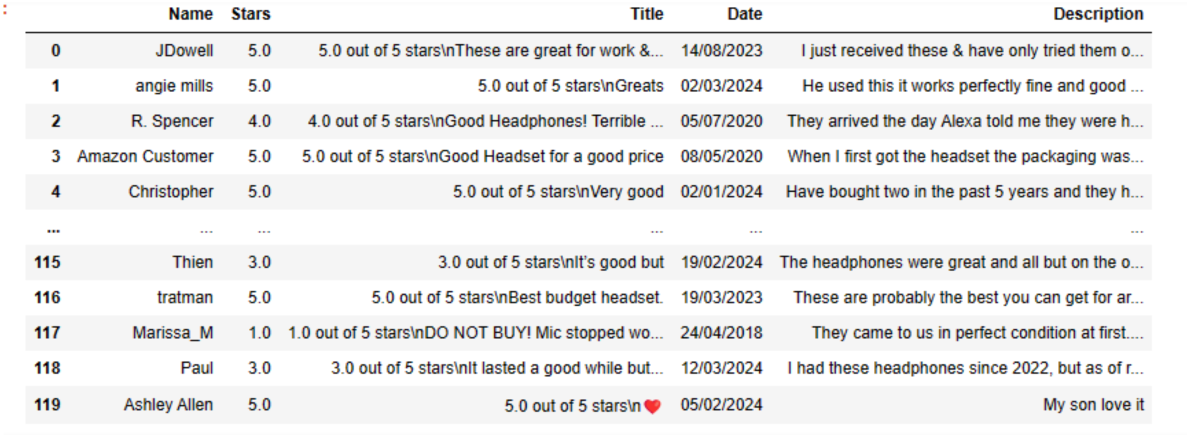
Kaarthikeyan Swaminathan
Well it is an IPhone!!!

Akhilesh K.
Last of the small phones

Abhimanyu Shrivastav

Fig. 5 Product Reviews of iPhone 13

The conducted market research shows the gain and pain points for the product from the customer reviews. In this section, the experimental results are gathered and discussed.

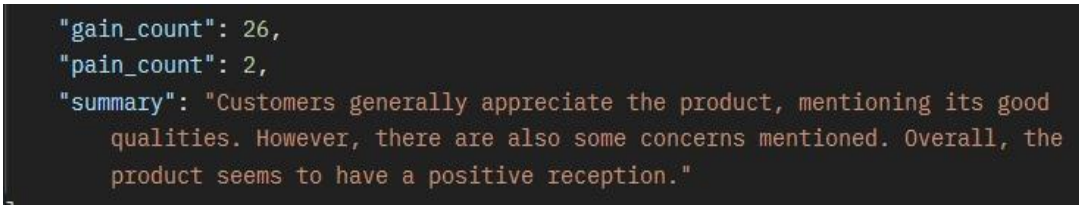


	Name	Stars	Title	Date	Description
0	JDowell	5.0	5.0 out of 5 stars\nThese are great for work &...	14/08/2023	I just received these & have only tried them o...
1	angie mills	5.0	5.0 out of 5 stars\nGreats	02/03/2024	He used this it works perfectly fine and good ...
2	R. Spencer	4.0	4.0 out of 5 stars\nGood Headphones! Terrible ...	05/07/2020	They arrived the day Alexa told me they were h...
3	Amazon Customer	5.0	5.0 out of 5 stars\nGood Headset for a good price	08/05/2020	When I first got the headset the packaging was...
4	Christopher	5.0	5.0 out of 5 stars\nVery good	02/01/2024	Have bought two in the past 5 years and they h...
...	...	...	...	...	...
115	Thien	3.0	3.0 out of 5 stars\nIt's good but	19/02/2024	The headphones were great and all but on the o...
116	tratman	5.0	5.0 out of 5 stars\nBest budget headset.	19/03/2023	These are probably the best you can get for ar...
117	Marissa_M	1.0	1.0 out of 5 stars\nDO NOT BUY! Mic stopped wo...	24/04/2018	They came to us in perfect condition at first....
118	Paul	3.0	3.0 out of 5 stars\nIt lasted a good while but...	12/03/2024	I had these headphones since 2022, but as of r...
119	Ashley Allen	5.0	5.0 out of 5 stars\n❤	05/02/2024	My son love it

Fig. 6 Extracted Reviews Constituting the Dataset

The dataset consists of 150 reviews extracted for headset product. The dataset is preprocessed to maintain only the required columns or features. The required feature set shall consist of the title of the review and review description. The Extracted reviews is shown in the Figure 6. To begin with, the reviews are fed to the Seq2Seq model as input. This shall be responsible for text summarization. Variable length input reviews are summarized as fixed-size context vector. Word by word, the recurrent layers process the input sequence, capturing details about the context and content of the review. This is then passed on to the LSTM model for sentiment analysis. Input sequences pass through an embedding layer, converting discrete word representations into continuous vector representations. Once trained, the model is ready for sentiment analysis on new, unlabeled product reviews by passing the review text through the model, which outputs the predicted sentiment label. For example, for the given reviews of a product, the summed gain points and pain points shown in the Figure 7 along with the overall summary of the product is achieved as:

{
"gain_count": 26,
 “**pain_count":** 2,
 "summary": "Customers generally appreciate the product, mentioning its good qualities. However, there are also some concerns mentioned. Overall, the product seems to have a positive reception."
}



```
"gain_count": 26,
"pain_count": 2,
"summary": "Customers generally appreciate the product, mentioning its good
    qualities. However, there are also some concerns mentioned. Overall, the
    product seems to have a positive reception."
```

Fig. 7 Gain Points, Pain Points and Summary for the Product

Here, the gain point for the product is 26 whereas the pain point for the product is 2. This means, on an average the customers have found more positive features in the product as compared to negative features. The ‘summary’ specifies how the customers have perceived the product. In-spite of some drawbacks in the product, many features are still liked by the customers.

C. Deployment

This project uses Docker containers and a cloud platform for reliable and scalable deployment. Docker ensures the application runs consistently across environments by packaging it with all its dependencies. A cloud platform like Vercel or Netlify simplifies deployment and manages tasks like scaling to handle changing demands. This approach allows for independent deployment and scaling of each microservice within the application.

V. SENTIMENT INTENSITY ANALYSER TOOL

By analyzing customer reviews and product details, we can provide valuable insights into the quality and perceived value of a product, helping customers make informed decisions about their purchases. Our idea is to code a program which uses python which takes input of user reviews and the details of a product and rates that particular product on various subjective parameters and gives a detailed review about the product. Finally, it gives an overall review of the product to the customers, pain points and gain points of the products. Our idea can be applied to any e-commerce product such gadgets, clothing, appliances, groceries and even vehicles too. Our concept would solve the product legitimacy problem and aid the customers in making a decision either to purchase the product or not. The stop words library can be very useful in sentiment analysis of reviews because it allows us to filter out common words that are unlikely to carry any sentiment. However, as mentioned earlier, it is important to make sure that negative words such as "not" are not removed from the review text. Removing stop words from text can help improve the accuracy and efficiency of text analysis algorithms by reducing the noise in the data.

The SentimentIntensityAnalyzer tool employs the VADER lexicon, which contains over 7,500 words, emoticons, and idioms each assigned a sentiment score ranging from -1 (most negative) to +1 (most positive), with modifiers like "very" considered. It analyzes text by breaking it into words, looking up each in the lexicon to aggregate and normalize scores into an overall sentiment score ranging from -1 (negative) to +1 (positive). This tool provides insights into the sentiment expressed in social media and other text, offering separate scores for positive, negative, and neutral sentiments based on the words' perceived emotional intensity.

A. Topic Modeling and Document Summarization

Keywords are identified from the gathered data using the Rapid Automatic Keyword Extraction (RAKE) method in order to conduct additional analysis and categorization. RAKE initially detects key phrases that serve as the foundation for sentiment analysis, summarizing documents, and modeling topics. By utilizing topic modeling and identifying key words that signify specific subjects, RAKE uncovers the underlying themes within product discussions, offering a more comprehensive understanding of product attributes. RAKE not only generates brief summaries of documents, but also provides concise summaries of product reviews by extracting important information and user opinions through LexRank or TextRank algorithms. Next, we are going to conduct sentiment analysis to determine the overall sentiment of reviews (whether they are positive, negative, or neutral) to be able to evaluate customer satisfaction. When used together, these techniques can assist us in determining the credibility of a product.

B. Clustering

The k-means algorithm categorizes each data point into a cluster based on its distance from the cluster's center, repeating this process until the clusters arrive at a stable state. One application of k-means clustering in mobile phones is to categorize phones based on their prices, reviews, and brands. This allows companies to understand market patterns and helps consumers make educated choices when buying products. When trying to find products, groups of similar items can be formed by focusing on specific criteria or factors that are significant in making purchasing choices. By utilizing a mobile phone dataset from Kaggle which includes information such as brand, cost, CPU, camera specs, battery longevity, and display size, clustering can streamline comparison and enable well-informed decision making that aligns with individual preferences and requirements.

The first clusters might be formed making use of the parameters or factors which are highly favored or deemed crucial by the users. For example, brand and cost are likely to be the most important factors for many consumers when buying a cell phone. The groupings could change as users provide feedback and their behavior is taken into account, in order to more accurately represent the interests and requirements of the user community. This could lead to a more customized and pertinent shopping experience for customers, leading to increased satisfaction and loyalty. Figures 8 and 9 display customer evaluations of mobile phones organized by price, rating, and brand name through the use of clustering and sentiment analysis. K-means clustering algorithms are able to group comparable data points, allowing for the identification of clusters of mobile phones based on their brand, rating, and price.

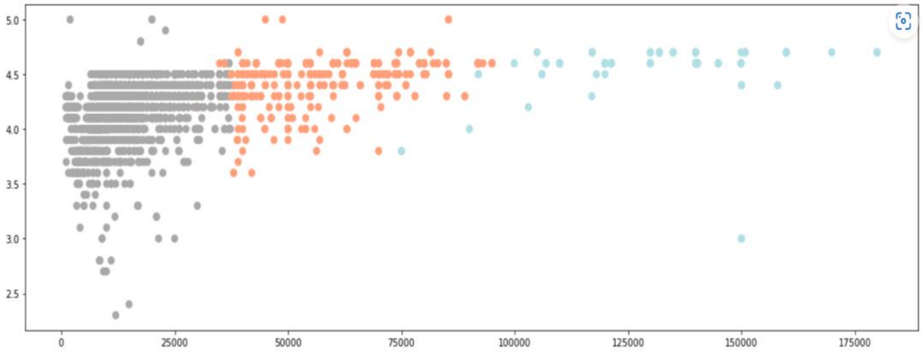


Fig. 8 Price Vs Rating

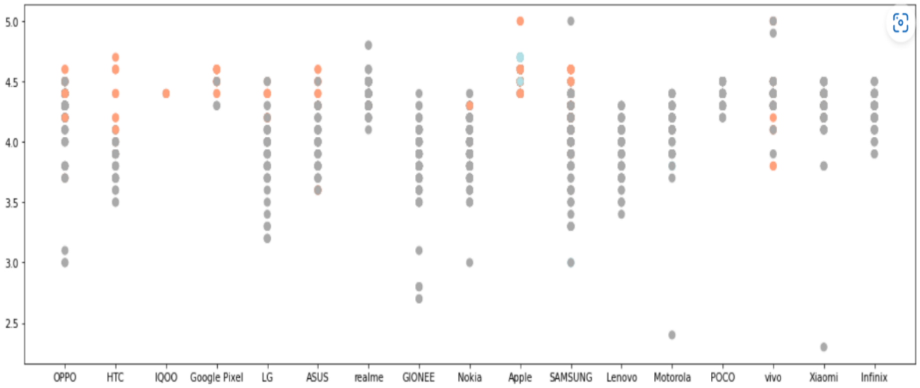


Fig. 9 Brand Vs Rating

Cluster analysis reveals several interesting facts about mobile phone market. Popular brand phones are usually more expensive but have wider ratings. It means consumers are prepared to shell out more for phones from familiar brands even if the phones are not the best performing or feature-packed. Budget-friendly phones can be good value for money, but they're also more likely to be from obscure brands. That means customers need to be careful when choosing budget phones - they might not be getting the same build quality or performance as a more expensive phone from a big name. Top names generally receive higher ratings on phones, although they also have a wider range of ratings indicating that customers are more lenient about minor defects in these products. Conversely, lesser-known brands may have competitive prices but not necessarily the brand recognition and support of larger brands. Smartphone failures prompt negative reviews, which highlight product quality and customer service. Its diagram illustrates how brand and rating influence consumer choices in the mobile phone market. Using clustering and sentiment analysis methods on customer reviews, businesses can gain valuable insights into consumer preferences and perceptions to inform strategic marketing and product development efforts. Blue bars show phone model prices; orange lines show average ratings; and shaded areas around them show the range of ratings - reflecting customer feedback variability. Analysis of Mobile Phone Models: Price vs. Customer Ratings shown in Figure 10.

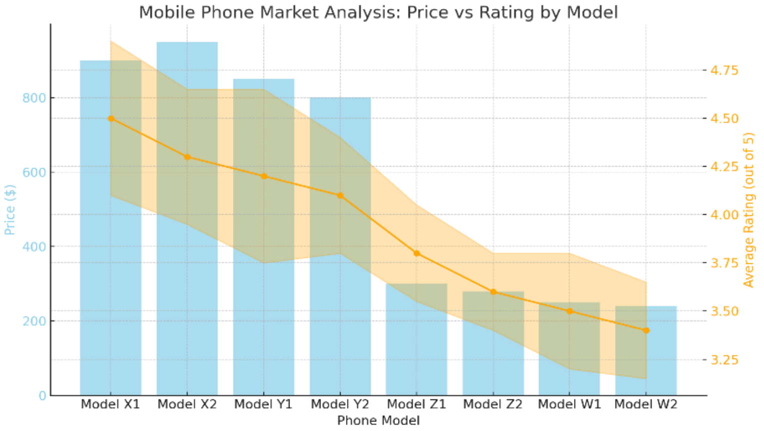


Fig. 10 Analysis of Mobile Phone Models: Price vs. Customer Ratings

Post-purchase rationalization explains why consumers justify expensive items by emphasizing their good qualities and minimizing their bad ones. This psychological mechanism helps with regret or cognitive dissonance related to expensive purchases. Positive reinforcement reduces discomfort with the high price, and shows how marketing and branding influence consumer perceptions. Company can capitalize on this by branding high-end mobile phones as premium products that reinforce perceptions of premium quality and experience. Such biases demonstrate how consumer evaluations might be biased towards products perceived as superior quality because they are marked with a price tag, in accordance with behavioral economics principles.

The sentiment intensity analyzer will calculate an average polarity score for each review and display it in its column along with a list of reviewed websites. Also, based on the product's positive, negative and neutral polarity scores, a custom message will be generated to help the user make a purchase decision. By aggregating mobile phones by price, brand, and ratings, we can see consumer preferences and trends in customer feedback

VI. DATA ANALYTICS IN ENHANCING CUSTOMER SATISFACTION

Data analytics are increasingly used in e-commerce and have shown to be a useful tool for improving customer satisfaction while shopping. AI-based personal shopping helps offline retail stores as well by combining artificial intelligence with retail shopping. This technology tries to provide clients a customised, easy to use and efficient purchasing experience. Usually, it is a personal chatbot assistant where the person specifies what they want in a product - price, color, material, location - and is shown a list of stores that stock the product to which they refer - along with reviews and ratings about the shops. Such technology can allow buyers to try-on a product before actually purchasing it. This saves you time &amp; effort while shopping. Additionally the system analyses customer and market distribution data for client preferences and trends. This helps retailers understand their audience and offer relevant items and services. That information helps them develop customized marketing strategies which will be received positively by clients. And it includes timing analysis of promotions to see when a company should advertise its product or introduce a new product. It may also analyze product data and make recommendations based on the client preferences. The software might recommend something based on something a customer has done before - past purchases, product reviews or even social media activity. From this point onwards, customers will get tailored suggestions according to their requirements and interests. The promotion timing function lets businesses time their promotions to maximize sales and customer satisfaction.

A. Geographical Location Exploration

Geographical location approach is a way of studying and analyzing phenomena, taking into account their geographical locations. It uses the physical and cultural features of a location to analyse how these affect the behaviours, attitudes and practises of its inhabitants. That means exploring human interactions with their environment - including climate, terrain, natural resources, and transport systems. We have considered market distribution, customer distribution and revenue distribution in our work.

B. Market Distribution

In terms of geographical location analysis, market distribution means reaching consumers in different geographic regions. This entails a detailed analysis of physical and cultural features such as terrain, climate, infrastructure, population density and cultural preferences to identify most appropriate distribution channels and strategies. So we first processed the geographic dataset by classifying zip codes by centroids and assigning associated longitude and latitude values based on those zip codes. It was necessary because one zip code prefix can include multiple latitude and longitude coordinates. We can analyze the geographic characteristics of each region and adapt our distribution strategies and product allocation so that goods and services reach consumers quickly and cheaply.

C. Revenue Distribution

Revenue distribution analysis studies how revenue is created and distributed across geographic regions. Often this analysis reveals important consumer behaviour / market demand / economic trends in particular areas. The distribution of product revenue was investigated by processing the order dataset. It started by removing duplicate data. Then we joined the order dataset with the geo dataset by joining the order ID column, since the location of each order was included in the geo dataset. That combined dataset enabled us to examine revenue distribution within geographic regions. We examined such things as the number of orders, average order value, and total revenue per region. Customer distribution: First, we processed the geographic data set. For each zip code prefix, we clustered longitude and latitude values based on their respective centroids because one zip code prefix may contain several unique latitude and longitude coordinates. This is the total number of customer IDs in the "count" column.

D. Customer Distribution

Firstly, we processed the geographical data set. We clustered zip codes and assigned longitude and latitude values based on their respective centroids, considering that a single zip code prefix might encompass multiple unique latitude and longitude coordinates. The "count" column represents the total number of customer IDs. Subsequently, we removed all duplicates from the geographical data set and generated a scatter mapbox plot to visualize customer distribution.

E. Promotion Timing

The timing and frequency of promotional activities used by businesses to attract customers and increase sales is referred to as promotion timing. Timing is an important factor in the success of promotional campaigns because it influences consumer interest and the promotion's effectiveness in achieving its goals. In our method we analysed to find the best hour, best day and best time. As the first step we changed the purchase time data type to datetime type.

a) Best hour: First we extracted the hour component from the purchase time using pandas accessor then we created a distribution plot to visualise the best hour.
b) Best day: First we extracted the day component from the purchase date using pandas accessor then we created a distribution plot to visualise the best day.
c) Best month: First we extracted the month component from the purchase date using pandas accessor then we created a distribution plot to visualise the best month.

F. Customer Behavioral Exploration

It is crucial to grasp customer behavior now more than ever, particularly with the rise of AI in purchasing assistance. To gain a deeper understanding of our customers' actions, we categorized them according to their length of time with us, how often they engage with our products or services, the amount they spend, and how recently they have interacted with us. This process led to the discovery of eight categories of clients, each possessing distinct traits and requirements. We have grouped customers according to their relationship with the store in order to gain a deeper insight into how they interact with our AI-powered shopping assistance. These clients appreciated the incentives and rewards program offered by our company, which encouraged repeat purchases from loyal customers. Presents from devoted clients were appreciated, and they thought that their patronage was highly regarded.

This data can be leveraged to deliver more personalised and effective AI-based shopping assistance to clients, ultimately increasing customer satisfaction and loyalty. In this article, we are going to look at the strategies and approaches employed during our study of consumer behavior, and the way we can make use of these insights to improve AI - driven shopping assistance

G. Virtual Try-on

Our algorithm creates a lifelike representation of the way the clothing is going to look on a 3D human model by matching their characteristics and preserving the texture of the clothing. Our technology could one day change the way people shop for clothes online, by providing them with the information they need to make an informed decision. Customers may have a more positive shopping experience and also be more satisfied with their purchases if they are able to visually see how a piece of clothing would look on a 3D model with the correct texture. This will ultimately result in a boost in sales for retailers.

Additionally, our method could greatly reduce time and cost required to produce top - notch 3D models for uses such as virtual fitting. Retailers can easily create precise 3D models of their products, eliminating the need for extensive manual editing. This simplifies and also reduces the cost of integrating virtual try-on features on their sites. This will lead to an improved shopping experience for customers, potentially increasing sales for stores.

H. Product Analysis/Exploration-popularity Analysis

By counting the number of reviews each Product received, we could figure out which category was most popular. The product with the highest number of reviews and a positive review rating was identified. Businesses are able to make use of this information to identify popular products that are highly sought after and may be used to increase sales. In general, exploring and analyzing products can offer important insights into the latest market trends and what consumers are looking for. By analyzing the popularity of their products and services, businesses are able to enhance their product offerings, boost customer satisfaction, and make necessary improvements. This kind of inquiry can help pinpoint possible opportunities for growth and improvement, leading to increased profits and success in the market.

VII. PRODUCT RECOMMENDATION SYSTEM

Integrating a customized product recommendation system is able to enhance the shopping experience for customers and provide a personalized element to their shopping journey. Customers are more inclined to buy the products suggested by the system because they are similar to the product they originally looked at, resulting in higher profits because of the company. Customer loyalty and satisfaction can be improved by making use of AI-based product recommendation systems, which can provide companies with useful insights into what their customers really want, and in turn, this information can be used to create better sales and marketing campaigns. Through carefully analyzing customer actions and preferences, companies are able to tailor product recommendations to meet specific needs and desires, ultimately enhancing the overall shopping experience for their customers.

VIII. CONCLUSION

In conclusion, AI data analytics improves the shopping experience by offering personalized recommendations and virtual try-on options, aiding customers in making informed choices. Our app sorts Amazon reviews based on their sentiment (positive, negative, or neutral) by analyzing keywords and sentiments. This approach allows us to grasp customer preferences and emotions, while also pinpointing areas of dissatisfaction or satisfaction - key aspects of products that impact consumer choices. Sophisticated NLP methods, like LSTM models used for sentiment analysis, will be integrated into product-oriented crawlers to gather and evaluate customer reviews. This comprehensive strategy provides businesses with guidance on improving product development and marketing in order to increase operational effectiveness and customer satisfaction.

References

[1] Abdo, A. A., Alhajri, K., Alyami, A., Alkhalaf, A., Allail, B., Alyami, E., & Baaqeel, H. (2023). AI-based Spam Detection Techniques for Online Social Networks: Challenges and Opportunities. *Journal of Internet Services and Information Security, 13*(3), 78-103.

[2] Boldi, P., Codenotti, B., Santini, M., & Vigna, S. (2017). UbiCrawler: a scalable fully distributed webcrawler, http://vigna.di.unimi.it

[3] Chauhan, C., & Sehgal, S. (2017). Sentiment analysis on product reviews. *In IEEE International Conference on Computing, Communication and Automation (ICCCA)*, 26-31. https://doi.org/10.1109/CCAA.2017.8229825

[4] Chauhan, C., Patel, P., & Chauhan, S. (2019). Sentiment analysis of e-commerce platforms: A case study on Amazon and Flipkart. *International Journal of Scientific Research in Computer Science and Engineering*, *7*(2), 60-66.

[5] Dey, M. K., Shamanta, D., Chowdhury, H. M. S., & Ahmed, K. E. U. (2010). Focused web crawling: a framework for crawling of country based financial data. *In IEEE 2nd IEEE International Conference on Information and Financial Engineering*, 409-412.

[6] Hati, D., Sahoo, B., & Kumar, A. (2010). Adaptive focused crawling based on link analysis. *In IEEE 2nd International Conference on Education Technology and Computer*, *4*, V4-455.

[7] Htay, S. S., & Fong, S. M. T. H. S. (2013). Linguistic features based on POS tags for extracting the essence of customer reviews. *International Journal of Computer Applications*, *80*(11), 25-31.

[8] Htay, S. S., & Lynn, K. T. (2013). Extracting product features and opinion words using pattern knowledge in customer reviews. *The Scientific World Journal*, *2013*(1), 394758. https://doi.org/10.1155/2013/394758

[9] Kamoonpuri, S. Z., & Sengar, A. (2023). Hi, May AI help you? An analysis of the barriers impeding the implementation and use of artificial intelligence-enabled virtual assistants in retail. *Journal of Retailing and Consumer Services*, *72*, 103258. https://doi.org/10.1016/j.jretconser.2023.103258

[10] Kausar, M. A., Dhaka, V. S., & Singh, S. K. (2013). Web crawler: a review. *International Journal of Computer Applications*, *63*(2), 31-36.

[11] Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies, 5*(1), 1-167.

[12] Mali, S., & Meshram, B. B. (2012). Implementation of multiuser personal web crawler. *In IEEE CSI Sixth International Conference on Software Engineering (CONSEG)*, 1-12.

[13] Muhammad, A. N., Bukhori, S., & Pandunata, P. (2019). Sentiment analysis of positive and negative of youtube comments using naïve bayes–support vector machine (nbsvm) classifier. *In IEEE International Conference on Computer Science, Information Technology, and Electrical Engineering (ICOMITEE)*, 199-205. https://doi.org/10.1109/ICOMITEE.2019.8920923.

[14] Muhammad, A. N., Khan, M. Z., & Ali, M. S. (2022). Sentiment analysis of YouTube comments using deep learning techniques. *International Journal of Advanced Computer Science and Applications, 13*(2), 45-51.

[15] Oleksandr, K., Viktoriya, G., Nataliia, A., Liliya, F., Oleh, O., & Maksym, M. (2024). Enhancing Economic Security through Digital Transformation in Investment Processes: Theoretical Perspectives and Methodological Approaches Integrating Environmental Sustainability. *Natural and Engineering Sciences, 9*(1), 26-45.

[16] Pan, X. Y., Chen, L., Yu, H., Zhao, Y. J., & Xiao, K. N. (2019). Survey on research of themed crawling technique. *Application Research of Computers*, *37*(5).

[17] Parvez, M. S., Nguyen, T. D., & Tran, S. H. (2021). Deep learning-based sentiment analysis for product review. *Journal of Information Technology Research*, *14*(1), 1-20.

[18] Parvez, M. S., Tasneem, K. S. A., Rajendra, S. S., & Bodke, K. R. (2018). Analysis of different web data extraction techniques. *In IEEE International Conference on Smart City and Emerging Technology (ICSCET)*, 1-7. https://doi.org/10.1109/ICSCET.2018.8537333.

[19] Pavalam, S. M., Raja, S. K., Jawahar, M., & Akorli, F. K. (2012). Web crawler in mobile systems. *International Journal* of *Machine Learning and Computing*, *2*(4), 531-534.

[20] Shah, S. A. A., Ahmed, N., & Asim, M. (2021). Bi-directional LSTM with CNN for e-commerce entity detection. *IEEE Access*, *9*, 66729-66739.

[21] Shah, S. A. A., Masood, M. A., & Yasin, A. (2022). Dark web: E-commerce information extraction based on name entity recognition using bidirectional-LSTM. *IEEE Access*, *10*, 99633-99645. https://doi.org/10.1109/ACCESS.2022.3206539.

[22] Sunarto, A., Kencana, P.N., Munadjat, B., Dewi, I.K., Abidin, A.Z., & Rahim, R. (2023). Application of Boosting Technique with C4. 5 Algorithm to Reduce the Classification Error Rate in Online Shoppers Purchasing Intention. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, 14*(2), 1-11.

[23] Yang, L., Chen, W., & Zhang, Z. (2020). Sentiment analysis model integrating sentiment lexicons and convolutional neural networks. *Knowledge-Based Systems*, *188*, 105034.

[24] Yang, L., Li, Y., Wang, J., & Sherratt, R. S. (2020). Sentiment analysis for E-commerce product reviews in Chinese based on sentiment lexicon and deep learning. *IEEE Access*, *8*, 23522-23530. https://doi.org/10.1109/ACCESS.2020.2969854

[25] Yemunarane, K., Chandramowleeswaran, G., Subramani, K., Ahmed, A., & Srinivas, G. (2024). Development and Management of E-Commerce Information Systems Using Edge Computing and Neural Networks. *Indian Journal of Information Sources and Services, 14*(2), 153–159. https://doi.org/10.51983/ijiss-2024.14.2.22

[26] Zhang, L., Wang, J., & Liu, L. (2021). A novel deep learning approach for sentiment analysis of product reviews. *Neurocomputing, 419*, 150-159.

[27] Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(4), e1253. https://doi.org/10.1002/widm.1253

[28] Zhang, Z., Guo, J., Zhang, H., Zhou, L., & Wang, M. (2022). Product selection based on sentiment analysis of online reviews: An intuitionistic fuzzy TODIM method. *Complex & Intelligent Systems*, *8*(4), 3349-3362.

[29] Zhang, Z., Yang, Y., & Deng, W. (2020). A sentiment analysis-based intuitionistic fuzzy TODIM method for product selection using online reviews. *IEEE Access*, *8*, 38908-38918.