International Journal of Information Technology and Computer Science Applications (IJITCSA)
p-ISSN: 2964-3139 e-ISSN: 2985-5330
Vol. 02, No. 02, page 91 - 98
Submitted 01/05/2024; Accepted 07/05/2024; Published 10/05/2024

Managing the E-commerce Data Deluge through Text Analytics and Web Management (Overview of Amazon.com)

Baru Khan Bau
Institute of Science and Technology, Tribhuvan University, Nepal
e-mail : barukhan.bau@gmail.com

Corresponding Autor: Baru Khan Bau

Abstract
Today, more than 80% of the big data handled in the e-commerce industry is text and unstructured data. Text analytics is an automated process for analyzing text and extracting useful information from it. It can discover trends and relationships in data. Web analytics is the collection, processing, and analysis of data in order to draw conclusions to optimize usability on a website. Web analytics can be used to improve the usability of a site by analyzing user behavior patterns such as time spent on the site, abandonment rates, most frequently accessed products, click-through rates, etc. It can also help analyze the interests of different user demographics, as it tracks granular details such as user demographics, age and gender, geography, and devices used as data. In order to obtain UpToDate information, the business can utilize business intelligence for real-time data processing, then they can practice stream analysis to analyse continuous flow of data. For instance, the business can collect instant information in Twitter or other social media and analyse it by using social media analysis. For website management, business can practice web analysis to analyse the customer's behaviours. Tracking the customer's activity, page view and conversion rate is important for business to analyse how to improve the website performance. Text analytics of comments received on Amazon can be used to group text data and produce results in terms of word frequency distribution and sentiment analysis. Text analytics could be used for decision making, improving service quality, and developing new business models.

Keywords: Text analytics, Web Analytics, Social Media, Stream analytics, Geospatial Analysis, Website Management

1 Introduction
By 2025, the total amount of data worldwide is projected to be 175 zettabytes [1]. How to manage and utilize such vast amounts of data is an important business strategy in the e-commerce industry. Today, Amazon.com and many other e-commerce organizations are analyzing big data to market more strategically to their customers and facilitate optimal decision making for them and their customers. According to Madnira [2], structured, semi-structured, and unstructured analysis with a variety of large data sets can lead to better and faster decision making. In this research paper, a study will be conducted on the application of text analytics in the e-commerce industry, including Amazon.com.

1.1 *Data Analytics on E-commerce*
According to Kittisak [3], unstructured data in the e-commerce industry are clicks, text, various signals, links, tweets, voice, images, audio, and video. Text analytics is an automated process for analyzing text and extracting useful information from it [2]. It analyzes social media posts such as Facebook and Twitter, review and comment sections, and other types of text and uses tools such as machine learning to find the meaning of that text. Text mining is the process of analyzing textual information to identify patterns. By analyzing customer behavior patterns on the site, trends in products purchased, reviews, etc., it is possible to provide the best offers, discounts, and information for each customer [3].

Sentiment analysis is a technique for text mining. It uses Natural Language Processing to automatically process and analyze unstructured text, such as social media posts, using AI to analyze sentence structure and grammar [4]. The analysis reads the sentiment of the text, separating and quantifying whether it is positive or negative. For example, the text of a product review is analyzed and given a number such as 0 for negative, 0.5 for neutral, and 1 for positive for each grammar and keyword to see what the final total number is as a numerical value. For e-commerce organizations such as Amazon, unstructured textual data, such as social media posts, and customer reviews can be mined to clarify customer satisfaction.

Web analytics is the collection, processing, and analysis of data in order to draw conclusions to optimize usability on a website. Web Analytics can be used to improve the usability of a site by analyzing user behavior patterns such as time spent on the site, abandonment rates, most frequently accessed products, click-through rates, etc. Web analytics can also help analyze the interests of different user demographics, as it tracks granular details such as user demographics, age and gender, geography, and devices used as data [5].

2 Problem Statement

In this digital age, most people are familiar with purchasing online, and e-commerce growth rates reached 17.1% in 2021 according to OBERLO [6]. The influence of lockdown during covid-19 restriction boosted the popularity of e-commerce industry more than before. The generalization of international e-commerce and high demand for such business makes the competition among the industry intense, and expectation for product quality and convenient website interaction high.

2.1 *Savage competition among multiple e-commerce sites*

The e-commerce market is highly competitive. The keys to getting ahead of competition are website performance and speed of changes on social media. In the research from Global Ranking [7], the Top 3 largest e-commerce companies are Amazon, followed by Alibaba and Maituan. These large businesses can spend on human resources, advertisement, and promotions to improve their web performance. Therefore, small businesses lose their customer base, and they are forced to lower the price to adjust to market price, leading to loss. Business needs to focus on storing big data and analysing it to get latest information continuously to survive among the competition.

2.2 *Customer expectation and behavior in e-commerce*

In the e-commerce industry, reputation is crucial as the customers do not rely on random websites as they cannot physically see and touch the products. And they expect to find what they are looking for easily with no stress when they interact with websites. There are also some customers who are still not sure what brand or product they want. In order to attract these customers to impulse buying, it is important to maintain web performance high. It is challenging to store and categorize vast amounts of data, and display and set the navigation to lead the customer smoothy to purchasing screen. They also must keep track of all geographical data of customers and operate efficient management to cover the international wide delivery. It is also important to analyse the millions of customer responses and product purchase reviews to improve the performance.

3 Proposed solution

In order to obtain up-to-date information, the business can utilise business intelligence for real-time data processing, then they can practice stream analysis to analyse continuous flow of data. For instance, the business can collect instant information in Twitter or other social media and analyse it by using social media analysis. By doing so, the business will not miss the changes in trend and record the competitor's information in the market. Therefore, it helps business to make decisions for business strategy, make faster moves than other competitors.

For website management, business can practice web analysis to analyse the customer's behaviours. Tracking the customer's activity, page view and conversion rate is important for business to analyse how to improve the website performance. Also, the customer responses and product reviews can be collected by text mining to structure the data to identify patterns, and sentiment analysis can identify additional insights such as customer's preference and new product ideas from the customer comment [8]. In terms of management, geospatial analysis is a must in e-commerce business as it helps to gather and display the geographical data and keep track of deliveries by GPS. The business can build efficient procedures to deliver the products.

3.1 *Data Integration*

Data integration is a process of combining data from multiple sources into a unified and single view [9]. It is important to integrate different kinds of data such as unstructured, structured, streaming, and batch data to store them

in a relational database that has structured column and map them to database entities. In this research we suggest integration of unstructured data, and analysis method to analyze e-commerce industry using stream analysis, text analysis, and social media analysis.

3.2 *Extraction, Transformation, and Load (ETL)*

One of the way of data integration is ETL. There are three process extraction, transformation, and load. The goal of extraction is to extract relevant data from multiple sources such as Twitter, Facebook, and Instagram as well as Amazon website to achieve objectives. Next phase is a transformation. In this phase, data extracted in previous phase is transformed to adoptable format to data warehouse by removing unneeded data, and applying normalization, standardization, and correction. The process of loading transformed data into data warehouse is known as load. However, when real-time data is procced ELT process is used instead of ETL and stored in data lake [9], [10].

3.3 *Privacy issue*

Privacy issues are very important when dealing with text data retrieval, since most of the text data obtained from documents, emails, chats, online comments, and text content on social media may contain the customer's personal information. In the case of e-commerce organizations, text data handles personal information such as address, phone number, e-mail address, account number, etc. that customers enter when purchasing products. If these text data are illegally accessed by a third party, they may be at risk of fraud or identity theft. Even if it is within a company, it should be protected by limiting privileges. Once compromised, the data is difficult to delete, resulting not only in customer identity theft, but also in loss of credibility for the organization.

Data protection laws have been negotiated in many countries around the world. According to Boehm et al. [11], 57% of customers have lost trust in organizations and 31% of customers no longer use the services of organizations that have suffered data breaches. Such privacy risks can be managed by using techniques such as fully homomorphic encryption, adversarial learning, and secure components, from data pre-processing to validation of results, all of which are included, and PET covering all phases of the model pipeline [12].

4 Advantages and Disadvantages of Using Text Data Retrieval

4.1 *Advantage of using text data retrieval*

a. Data extraction at scale - By using text data retrieval, E-commerce organizations can quickly collect data on a massive scale. If text data retrieval is not used, organizations must manually filter through the sources one by one and collect data.
b. Cost and time efficient - It is cost and time efficient to extract data by text data retrieval rather than building complex system. It is automatic repetitive scraping that can replace hiring extra employees, so that the business can save the cost. For example, e-commerce businesses can use text scraping to extract customer reviews on the website rather than having surveys which are time consuming.
c. Data accuracy - One of advantages of retrieving text data by text data scraping is not only speed but also accuracy. It would be more likely to happen human error if collecting text data manually and that will effect on results and decision making later. It also happens error at least in small proportions, but it is easy to fix problems.

4.2 *Advantage of using text data retrieval*

a. Scraping can get blocked - When collecting text data from an external website, if the HTML structure of that website is changed, it may no longer be possible to collect text data. In such cases, it is necessary to change the program or scraping destination, which is time-consuming. Also, if scraping is performed too frequently, access may be denied based on log information.
b. Difficulty of learning the text scraping - Text scraping is convenient, but the employee needs to be trained to use all the coding. Even though the operation is more user-friendly nowadays, some employees are not motivated to learn logical coding. Training might need to be done, which will lead to increased cost and time consuming for the business.
c. Text data scraping itself does not perform the analysis - Before performing text data scraping, it is important to determine what data is needed in your objective, which means it is required to learn about analytics skill before extracting data unless you can understand it.

5 Dashboard

We have extracted the data source from Kaggle.com in order to analyze insights of the Amazon E-commerce performance in United State and build the dashboard shown in Figure 1. The datasets we used are following:

a. US E-commerce records 2020.csv
b. Test.csv
c. Train.csv

which are listed in the reference on last page. The visualization shown in dashboard can help to solve the problem statement mentioned in the previous section.

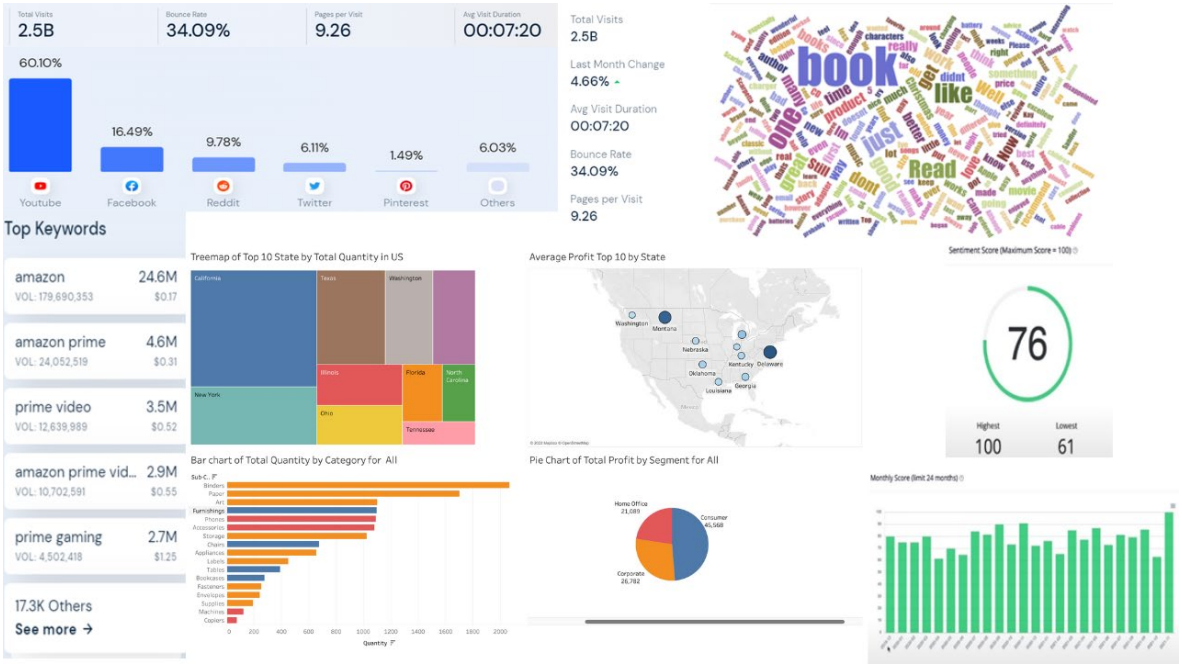

Figure 1. Dashboard on Amazon.com website

Sentiment analysis separates customer reviews word by word, and customer reviews are considered positive as they approach 100% and negative as they approach 0%. The highest sentiment score written on Amazon.com is 100%, and the lowest is 61%. The average, at 76%, means that customers are very satisfied with their product purchase, hence the positive reviews. The Word cloud in the dashboard displays the most frequently used words in the reviews in a large font. This shows that keywords related to reading, such as “book” and “read”, are observed. This may be due to the fact that Amazon was a marketplace for online book sales when it was founded, and still offers its own e-book description, such as kindle unlimited, as well as a wide selection of books.

As for the search keyword that reached Amazon.com, "amazon" was by far the most popular with 24.6 million. It can be estimated that users reach out to amazon.com and then search for the products they want within the amazon.com search engine. In addition, the high number of searches related to subscriptions offered by amazon, with 4.6 million for "amazon prime" and 3.5 million for "prime video", indicates that amazon prime is attracting a lot of interest. As shown in the bar graph on the dashboard, most of Amazon.com's social media traffic comes from Youtube, followed by Facebook and Reddit. Pinterest is low at 1%, which means there may be an opportunity to gain new customers in the future. The average time a user spends on a page is 7 minutes and 20 seconds with 9.26 pages per visit. Pages per visit is the average number of web pages viewed on a single website by a user or group of users. It is usually calculated by dividing the total number of page views by the total number of visitors. Littledata, which analyzed 3,698 e-commerce sites, found that the average number of pages per session ranged between 1.8 and 4.4. Comparing with these values, Amazon.com's average number of pages per session of 9.26 is well above the average, indicating very good usability [13].

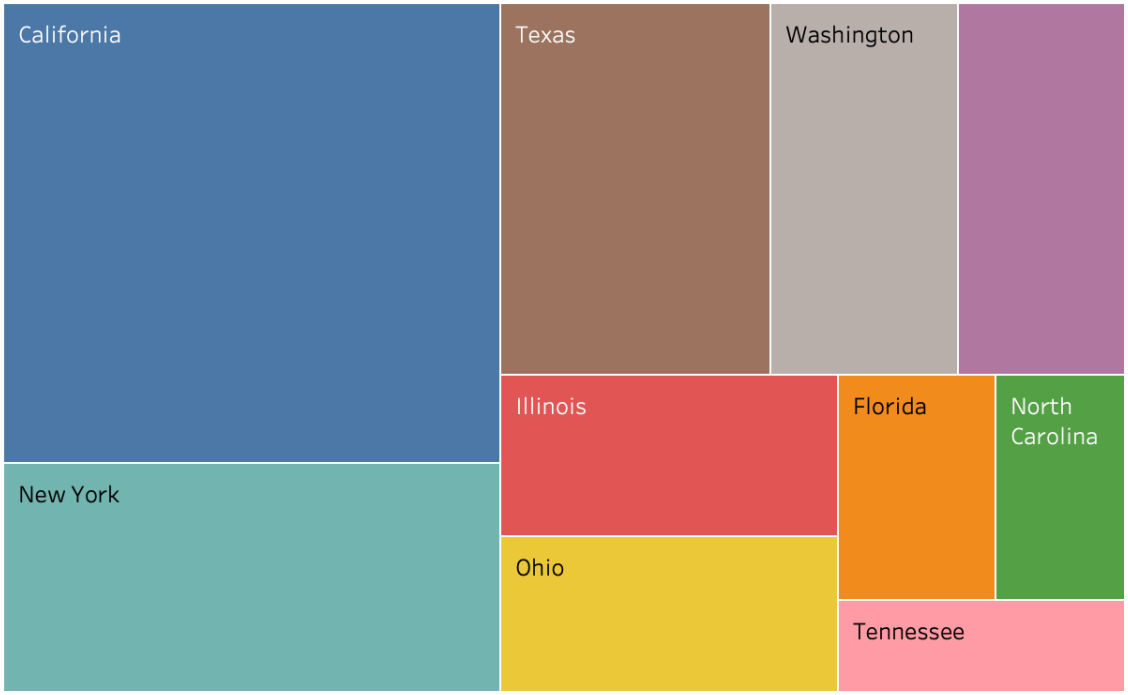

Figure 2. Frequency of data by state in E-commerce market in US

The tree map shows states in US, which is California, New York Texas, Washington, Pennsylvania, Illinois, Ohio, Florida, North Carlina, and Tennessee. The top states are ranked by total order quantity and the bigger size of panel means more frequent state in data set. The most frequent state is California which is 2633 out of 12476. Also, the panel can work as a filter. By clicking a panel of state, we can select the state and see the bar chart of total quantity by category and pic chart of total profit by segment for selected state.

We assume that a factor related to the frequency for each state is population. According to the US department of Commerce, the resident population of California is 39,538,223 which is the biggest number among all states. Also, the second and third most frequent states, New York and Texas which are 1,319 and 1,158 have a large number of resident populations which is 20,201,249 and 29,145,505.

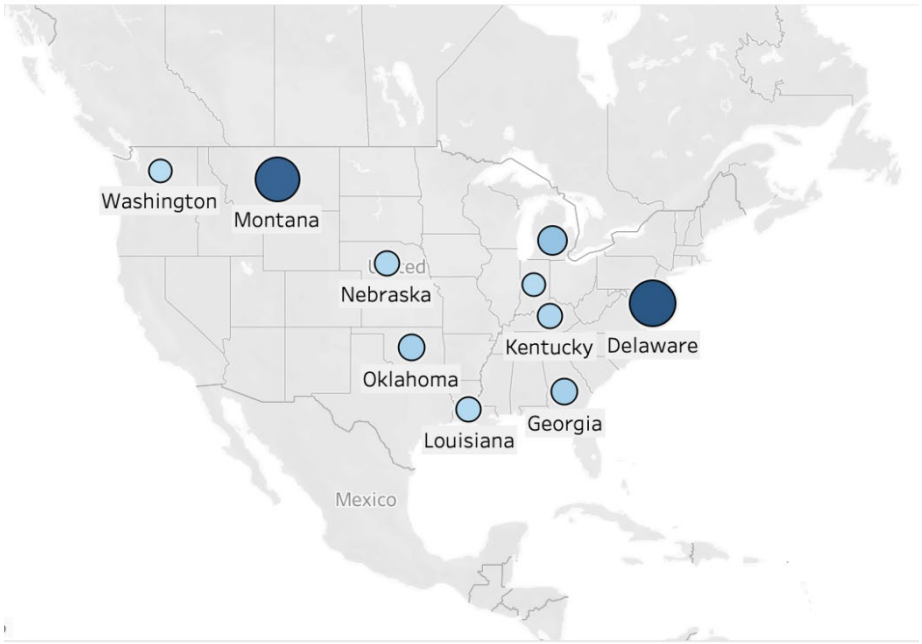

Figure 3. Top 10 state in E-commerce in US

In terms of geospatial analysis, a geo map shows Top 10 average profit by state. There are 10 states which is Delaware, Montana, Michigan, Georgia, Oklahoma, Kentucky, Nebraska, Louisiana, Indiana, and Washington. The size and darkness of the circle express the amount of profit. The bigger circle and dark bule means higher average profit. Analyzing the average of profit by state is important information to make a decision what states are more profitable for the organization.

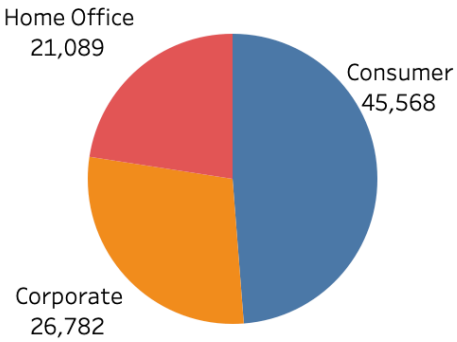

Figure 4. Different segments in E-commerce market in US

In the pie chart, there are three different segments which are Consumer, Corporate and Home Office. It is important to analyse the profit in each segment as profit is generally correlating with demand in the market. In the graph, Customer indicates the amount customer purchased online for personal use (B2C) whereas Corporate indicates the amount business purchased online for business use (B2B). Home Office means that the amount customer purchased online for business use which is mix of B2B and B2C. By analysing this graph, we can notice that Customer segment takes almost half proportion of overall (45.568%). It means that the Customer segment is most profitable in e-commerce, and we can assume market demands for Amazon e-commerce is occupied by B2C segment, followed by Corporate (26.782%) and Home Office (21.809%).

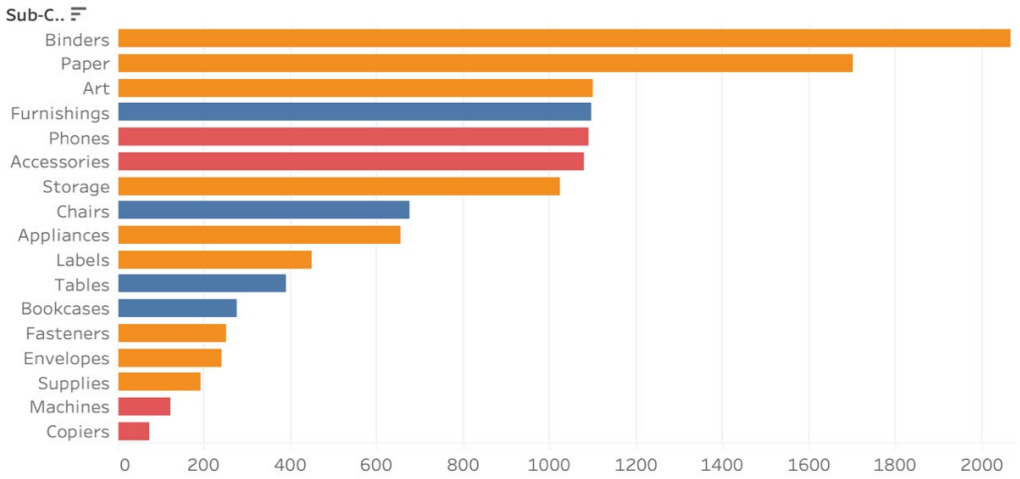

Figure 5. Order quantity in each category in E-commerce market in the US

The bar chart in Figure 5 shows the total order quantity in each category in Amazon e-commerce. From the bar chart, we analyse which categories have high demand and which does not. The colour on the bar chart indicates the major category types. For example, binder, paper, and art, and so on are considered as Office Supplies. We can notice that the Office Supplies category has high demand compared to Furniture and Technology. We assume this is because Office Supplies such binders and paper are relatively small objects and non-expensive goods. However, customers might find it resistant to purchase Furniture and Technology categories as they would want to physically see and touch before purchasing these goods and it is more expensive compared to Office Supplies.

5.1 *Challenges in Implementation*

Problems are likely to occur when data mining multilingual text. Currently, data mining uses algorithms and techniques independently for each language to support multilingual texts. Since many e-commerce sites, such as Amazon, are used by customers who purchase from many different regions and countries, the text obtained on e-commerce is multilingual. However, since different algorithms are used for each language, compatibility is not perfect, and existing text mining techniques and tools do not support multilingual documents, which sometimes causes many problems in the knowledge discovery and decision-making process. In particular, words with broad meanings, similarities, or the same spelling but with different meanings are still an open problem today.

5.2 *Data Collection Process on Web Crawling*

The data collection process begins with performing web scraping on e-commence web site such as Amazon web site and social media such as Twitter, Facebook, and Instagram including review of product of Amazon. A web crawler is a system for downloading a lot of pages from a website. One of uses of web crawler is web data extraction and web data mining. Web crawlers can go through serval new web pages by following the linked structure of the web page. Also, by using the graphical structure of the web pages, it can move from page to page, and it is designed to obtain web pages and put it into local repository.

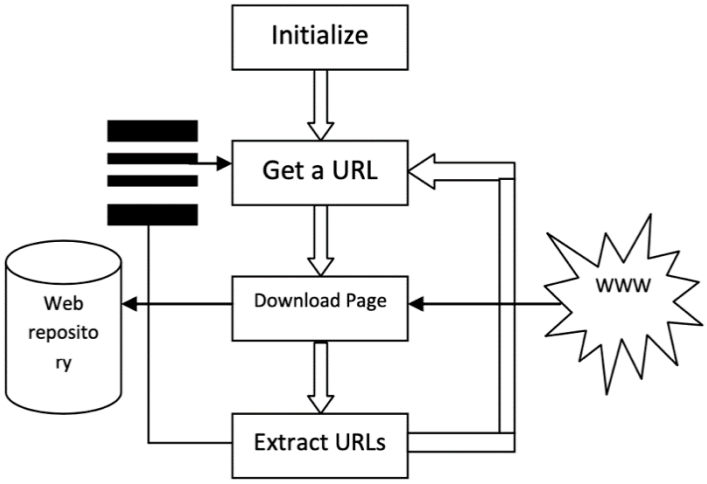

Figure 6. Flow chart of Web Crawler

5.3 *Data Cleaning*

Data cleaning is the process of removing or fixing inaccurate, unnecessary, and unformatted data in the dataset. It is an important process as the analysis will be unreliable if the data is not incorrect in the first place. For instance, data in the dataset must be stored in same format, otherwise it will lead to an error and some data might get lost when it's processed to next stage. The business should have a rule or template to follow when they extract data and practice data cleaning, so they know how to format and organize the dataset by following the normalized instruction.

5.4 *Data Mining*

Data mining uses techniques such as clustering and logistic regression analysis to discover and group patterns in data. Clustering is a technique for classifying data based on similarity. This is useful for grouping customers with similar behavior based on purchasing data and implementing different marketing initiatives for each group. Logistic regression analysis is a technique suitable for analyzing items for which "Yes" and "No" can be clearly defined. For example, it can be used to predict whether customers who have written reviews in the past will purchase a product when a product discount campaign is implemented. Artificial intelligence is sometimes used in data mining, and programming languages such as Python and R are often used in AI-based data analysis. In particular, Python is easy to use with its extensive libraries for data analysis and is an effective method for knowledge discovery to find patterns and relationships in data.

5.5 Data Visualization

There is a need to visualize bulk amount of data in understandable and accessible way while organizations generate various type of data. So that, the amount of available data on web sites has been increasing dramatically. It is not easy to handle and visualize massive amounts of data for the majority of users and visualizing skill is required

and more important to research. Data visualization can provide effective representation of data decision makers. Therefore, they can see analytics of their business and make a decision based on the visualization. For example, Word cloud is one of visualization techniques commonly used that allow users to obtain the content of a large amount of textual data. The significant terms could be automatically extracted and processed from a content. Also, the worlds are visualized with different font size and colors, so that it allows users to get significant terms in content instead of reading many description and customer reviews as shown in figure.

6 Conclusion

In conclusion, the business can solve their difficulties in their problem statement by implementing technical architecture mentioned in above and practicing various analysis. In stream analysis, the continuous flow data can be analyzed, and it is essential as keys of the e-commerce industry are the speed and up-to-date information. Text analysis can sort and group the text data to structure the meaningful information according to their objective. Amazon can utilize text analysis when they want to analyze the trend in customer preferences and keywords which might be customized for search engines, it provides the insights and ideas of business strategy and supports the decision making. They also can practice web analysis as the website platform performance is crucial when it comes to e-commerce. By analyzing the visitor's activity on the website such as conversion rate, they can build the improvement process to drive more online sales as same as social media analysis. The visualization of analytics aids to present the result of analytics as it is an easy and quick way to understand the overall analytics, making it accessible to others.

BIBLIOGRAPHY

[1]. "Ai and the global 'Datasphere': How much information will humanity have by 2025?," Data Universe, https://www.datauniverseevent.com/en-us/blog/general/AI-and-the-Global-Datasphere-How-Much-Information-Will-Humanity-Have-By-2025.html (accessed March 1, 2024).

[2]. S. Madnira, "Business intelligence (BI) approach for traffic accidents analysis," International Journal of Information Technology and Computer Science Applications, vol. 1, no. 2, Jun. 2023. doi:10.58776/ijitcsa.v1i2.32.

[3]. A. Kittisak, "Challenges and strategies for inventory management in small and medium-sized Cosmetic Enterprises: A Review," International Journal of Information Technology and Computer Science Applications, vol. 1, no. 2, Jun. 2023. doi:10.58776/ijitcsa.v1i2.30.

[4]. H. Xu, "Unimodal sentiment analysis," Multi-Modal Sentiment Analysis, pp. 135–177, 2023. doi:10.1007/978-981-99-5776-7_4.

[5]. Shujana, "Navigating healthcare challenges text analytics, data integration, and decision-making in the COVID-19 ERA," International Journal of Information Technology and Computer Science Applications, vol. 2, no. 1, pp. 54–62, Jan. 2024. doi:10.58776/ijitcsa.v2i1.123.

[6]. J. Gaubys, "How many people have smartphones? [Mar 2024 update]," Oberlo, https://www.oberlo.com/statistics/how-many-people-have-smartphones (accessed March 1, 2024).

[7]. "Largest e-commerce companies by market cap," CompaniesMarketCap.com - companies ranked by market capitalization, https://companiesmarketcap.com/e-commerce/largest-e-commerce-companies-by-market-cap/ (accessed March 1, 2024).

[8]. Amalia Nur Soliha, Tb Ai Munandar, and Muhammad Yasir, "Sentiment analysis of the use of digital banking service applications on Google Play store reviews using naïve Bayes method," International Journal of Information Technology and Computer Science Applications, vol. 1, no. 3, pp. 129–137, Sep. 2023. doi:10.58776/ijitcsa.v1i3.40.

[9]. F. Noor, "Development of a process for migration of data from relational to non-relational database," International Journal of Psychosocial Rehabilitation, vol. 23, no. 4, pp. 1261–1272, Dec. 2019. doi:10.37200/ijpr/v23i4/pr190452.

[10]. D. Seenivasan, "ETL (extract, transform, load) best practices," International Journal of Computer Trends and Technology, vol. 71, no. 1, pp. 40–44, Jan. 2023. doi:10.14445/22312803/ijctt-v71i1p106.

[11]. J. Boehm, L. Grennan, A. Singla, and K. Smaje, "Why Digital Trust Truly Matters," McKinsey & Company, https://www.mckinsey.com/capabilities/quantumblack/our-insights/why-digital-trust-truly-matters (accessed March 1, 2024).

[12]. S. Z. El Mestari, G. Lenzini, and H. Demirci, "Preserving data privacy in machine learning systems," Computers & Security, vol. 137, p. 103605, Feb. 2024. doi:10.1016/j.cose.2023.103605.

[13]. Littledata, https://lp.littledata.io/average/pages-per-session-(all-devices) (accessed May 1, 2024).