

Website: ijetms.in Issue:1, Volume No.4, January-2020 DOI:10.46647/ijetms.2020.v04i01.003

EDA(Exploratory Data Analysis) for Amazon Product Alexa

¹Ayesha Naureen, ²Dr. Satish Narayanrao Gujar

¹Research Scholar, Department of Computer Science and Engineering, University of Technology, Jaipur, India ² Supervisor, Professor, Department of Computer Science and Engineering, University of Technology, Jaipur, India ¹nooriekhan005@gmail.com, ²satishgujar@gmail.com

Abstract—In every Data Analysis or Data Science endeavour, EDA is a critical step. EDA is the investigation of a dataset to find patterns and anomalies (outliers), as well as to create hypotheses based on our knowledge of the dataset. Exploratory data analysis is an important part of every research project. It also gives tools for developing hypotheses by visualising and comprehending data, which is commonly done through graphical representation. I used the following EDA in this project. Techniques have been examined, and they will be used in the instance under discussion. Exploratory data analysis is useful for determining the underlying substance of data and for experimenting with statistical analysis in the form of descriptive and inferential analysis. Amazon alexa dataset has used in this paper and used exploratory data analysis to understand the features and underlying content in the dataset. From this, we used deep dive individual analysis on reviews given by users on an individual entity, and we were able to get the frequent topics and create word cloud visualisations, Term frequency, emotion word cloud, Circular dendogram, hierarchal dendogram, word association, polarity on negative and positive words.

Keywords—Machine Learning, EDA techniques

Introduction (*Heading 1*)

Exploratory data analysis (EDA) in statistics aids data analysts in gaining a visual understanding of the data's major properties, which leads to the formulation of hypotheses and the implementation of new experiments. There are two types of exploratory data analysis. To begin, there are two types of methods: non-graphical and graphical. Second, there are two types of methods: univariate and multivariate. The Alexa dataset is on reviews of the consumers of amazon product and this study uses EDA to comprehend it. Dataset from Alexa https://www.amazon.in/Echo-Dot-3rd-Gen improved/ product-reviews/B07W95FDFT/ref=cm cr dp d show all btm?showViewpoints=1&page=1&ref=cm cr dp d show all btm?showViewpoints=1&page=1&ref=cm cr dp d show all The total number of user reviews. We utilised topic modelling to construct topics on the text for a single entity, and R was used as the programming language. We also conducted a customer feedback search using the Google search engine for the Amazon product Alexa, which was done in R. We obtained the findings of the sentiment analysis of the product Alexa using R; emotions include pleasure, rage, happiness, trust, and so on.

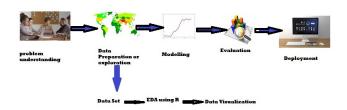


Fig 1: EDA Process

I. LITERATURE REVIEW

Roger D. Peng's exploratory data analysis 2012[1] This book contains a thorough examination of EDA. Benjamin Bengfort's has Apply Text Analysis along with Python that is Enabling Language Conscious Data Product by Machine Learning 2017[2]. In this study, the chapters on text preparation and wrangling (Chapter 3) and clustering for text similarity (Chapter 6) were applied. Allen B. Downey's Deliberate Stats: EDAThis book [3] addresses the complete process of data analytics, from data gathering through statistical outcomes. Nabiha Asghar's Yelp Dataset Challenge: Review Rating Prediction 2016[4] wrote about predicting review ratings. "The Philosophy of Exploratory Data Analysis," by I. J. Good, 1983[5]: - This work attempts to explain EDA in philosophical terms. Topic modelling is a technique for analysing text that has not been categorised. A overview of text mining topic modelling (2015) What are the different approaches in topic modelling and how are they used? [6] paper, authors present. The LDA Topic Model as well Word Co-occurrence are used to compute text similarity. In this research, Minglai Shao plus Liangxi Qin 2014 [7] created a text resemblance computation technique based on hidden themes models and word occurrences. Using Latent Dirichlet Allocation to Find the Natural Number of Topics: Some Observations According to a 2010 [8] report, https://www.yelp.com/dataset/challenge [9]. The document [Four Experiments on the Perception of Bar Charts - Tableau Research 2014 [10] discusses extensively with bar plots. Principles, promises, in addition problems of a data analytics within EDA and testing Data mining approaches and machine learning in a electronic design automation as well test are discussed within a study



Website: ijetms.in Issue:1, Volume No.4, January-2020 DOI:10.46647/ijetms.2020.v04i01.003

published in 2016 [11]. The paper The application of exploratory data analysis in auditing 2014 [12] demonstrates how EDA is used in auditing. Word cloud surveyor: text analytics based on word cloud (2014) [13] describes how to visualise text using a word cloud and their programme, word cloud explorer. Multiple text documents are seen as a word cloud. [14] 2015 This study is about using a word cloud to analyse numerous texts. Kushal Dave, Steve Lawrence, as well David M.P. Ennock (Kushal Dave, Steve Lawrence, and David M.P. Ennock, 2003) [15], had devised a system for automatically discriminating between good and negative reviews, using methods such as SVM with n-grams and metrics to assess performance (Precision and recall). Bo Pang, Lee, and Sriva Kumar (Bo Pang, Lee, and Sriva Kumar, 2002) [16], which was categorised not by subject but by general mood.

Data setDataset is in CSV format by Amazon Product Alexa.

	В	С	D	E	F
	Title	Rating	Text	Date	Туре
1	Good speaker, Bad s	3	This is very good har	2018-10-20	Verified Purchase
2	Very poor sound reco	1	It definitely has bette	2018-10-19	Verified Purchase
3	My first Smart Speak	5	This was launched o	2018-10-15	Verified Purchase
-4	Stunner	4	It's totally Palsa wasc	2018-10-15	Verified Purchase
5	it's amazing just read	5	Such an amazing p	2018-12-22	Verified Purchase
6	Worst software. I'm to	- 1	Worst software!!! Not	2018-11-19	Verified Purchase
7	A real competition to	4	At first i was confused	2019-02-19	Verified Purchase
8	Finally, a smart "spec	5	I got it on sale for 3k.	2018-10-26	Verified Purchase
9	Good music player b	2	Rate of speech cann	2018-10-21	Verified Purchase
10	it does not support fir	1	This device does not	2018-10-15	Verified Purchase
11	Nice & Decent Produ	5	I have already eco d	2019-10-22	Style Name: Echo
12	Cons	1	1.we have to talk nea	2018-10-28	Verified Purchase
13	Best at what it does.	5	Presently there are to	2018-11-13	Verified Purchase
14	it's a worth buying	5	I bought 3rd gen Ech	2018-11-10	Verified Purchase
15	DIFFICULT TO INST	4	BEFORE PURCHASE	2018-11-03	Verified Purchase
16	Waste product		Waste item,nothing f	2018-12-02	Verified Purchase
17	Just awesomeam lo	5	3rd gen product is ju	2018-10-20	Verified Purchase
18	Great Functionality,	1	Hi The functionality	2018-11-02	Verified Purchase
19	Bang for buck!	5	Don't know if there's	2018-10-19	Verified Purchase
20	Bought It once, WIII	5	1) Firstly, you may be	2018-11-07	Verified Purchase
21	Alexa is smart and in	5	So I got the 3rd gene	2018-10-16	Verified Purchase
22	Beauty with beast fer	5	It's jus beyond aweso	2019-01-21	Verified Purchase
23	It is only a good han-	1	Echo dot is a hands	2018-12-05	Verified Purchase
24	Google is infinitely b	1	Even the best device	2019-10-04	Style Name: Echo
25	Amazon's entry-level	4	Verdict-While the pre	2019-07-21	Verified Purchase
26	Disappointed due to	3	Its really made the n	2018-11-04	Verified Purchase
27	It's as a good as the t	5	An incredible techno	2018-10-21	Verified Purchase
28	Not up to the expect	3	Cons:1. Cant Line in	2018-10-24	Verified Purchase
29	A BIG NO	1	I am writing about m	2019-05-20	Verified Purchase
30	Great Product!!	5	I had been using the	2018-10-31	Verified Purchase
31	Alexa is better than I	5	It's true that Alexa co	2019-04-15	Verified Purchase
32	Cumbersome and no	1	Why the single star	2019-02-19	Verified Purchase

Fig 2: Sample dataset of Amazon Product Alexa

3.1 Problem Defination

People generally dislike looking at a column of numbers in a dataset to determine statistics; looking at numbers can be difficult and time consuming, not to mention unfriendly; the yelp dataset has many tables, so EDA was used to solve this problem; business entities want to know what reviewers are talking about, so a word cloud was used; moreover, most businesses want to know what's trending, so Topic modelling was used on a single entity after exploratory data; and most businesses want to know what'

Exploratory data analysis: Bar charts present numerical data in a highly particular style.

3.2 R is an open source statistical programming language. The R tool is utilised to discover the most significant characteristic in this study[16]. Amazon Reviews[17] data sets are accessed and analysed using R studio.

R is a data science programming language that is both dynamic and adaptable. This article explains how to classify data in R. In general, classifiers in R are used to predict certain category-related information, such as reviews or ratings of excellent, best, or worst.

SVM, Nave Bayes, Random Forest, and other classifiers are examples.

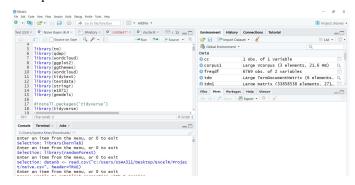


Fig 3: R studio IDE

It is necessary to authenticate and link R Studio to Amazon in order to retrieve Amazon reviews. After you've completed the authentication, you may move on to the next stage.

- **b)** Installing R Packages: In order to complete the analysis, you must first install the relevant packages. The package includes a number of functions that will be useful in assessing feelings.
- **d) Amazon review extraction**: It is to gather information from reviews on slightly topic using the hash tag "#."
- **d) Data Pre-processing besides Data Cleaning:** Unwanted phrases and words are removed from the data.
- e) Data Modeling and Transformations: Data is transmuted and organized in clear organised way to obtain emotions after retrieval and cleaning.
 - f) Sentiment Retrieval: Sentiment analysis is carried out.
- **g) Graphical Representation:** It's last step, wherever sentiments be plotted as well visualized by graph then word cloud.

Install the sentiment analysis packages that you'll need. The following are some of the packages that were used:

Tm-R is a text mining software written in R.

The R package Qdap (Quantitative Discourse Analysis Package) is developed to help with quantitative discourse analysis. The software acts as a link between qualitative discussion transcripts and statistical analysis and visualisation.

To analyse texts and instantly visualise keywords as a word cloud, use Wordcloud.

ggplot2- allows you to create a variety of graphs.

ggthemes - Using the ggthemes package's functions, modify the theme of a ggplot2 graph.



Website: ijetms.in Issue:1, Volume No.4, January-2020 DOI:10.46647/ijetms.2020.v04i01.003

tidytext — allows text to be converted to and from tidy formats, as well as easy switching between tidy tools and existing text mining software.

Textdata - Text analytics is commonly used to produce graphs, tables, and other visual representations.

stringr - to use wrappers

SnowballC - stemming is process of a reducing word to their root or base method.

RColorBrewer - is a tool for creating colour palettes that may be used in a variety of plots[18].

Term frequency

[19] observed that in the emotion mining classification job, existence of words is additional essential than its frequency, and in topic relevance classification task, the converse is true. It indicates that binary-value feature vector, in which entries specify whether word appears (value 1) or non (value 0), are additional effective used for review of polarity classification rather than a real-value feature vectors, with which entry value increases as term's frequency increases. We've used frequency to represent terms in this case. It allows us to count the most common term in a corpus by first converting our individual list of tokenized reviews into a complete list of iterative tokens that contains all reviews together, and then passing all words objects to the most common function to acquire common words. The findings of the phrase frequency analysis strongly corroborate the general favorable tone of the reviews. Terms like "play, just, song, time, use one, also, better, like, fantastic, home, smart, voice, device, music, can quality, sound product, good" indicate that Alexa is utilized in a favorable setting.in below fig it shown the term frequency of amazon alexa.

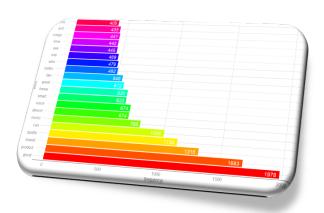


Fig 4: Term Frequency of Amazon Product Alexa

4.2 Word Associations

Correlation is a statistical approach that may show whether or not two variables are connected and how strongly they are related. This approach may be used to determine which words appear most frequently in conjunction with the most commonly occurring terms in survey replies, allowing for a better understanding of the context around these words.

Many reviewers link this device to the Amazon Echo thanks to word correlations with Alexa.in below fig it shown the word association of amazon product alexa.



Fig 5: Word association of Amazon Product Alexa

Product' is associated along with approving words alike amazing and good.

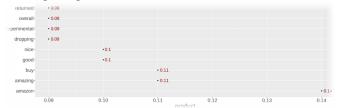


Fig 6: Word association of word "Product" of amazon product alexa.

4.3 Emotion Word Cloud

Emotions remain complicated part of a human survival. These feelings has impact on human decision-making as well help us communicate more effectively with break of world.process of recognizing a person's diverse sensations otherwise emotions is identified as emotion recognition or else emotion recognition (for example, joy, sadness, or fury). For the past few years, researchers working hard to systematize emotion identification. Some bodily activity, such as pulse rate, hand shaking, perspiration, and voice tone, can also reveal an individual's emotive state (Kratzwald et al. 2018), however emotion identification from text is difficult. Furthermore, with each passing day, new ambiguities and slang or terminology are added, making emotion identification from text added difficult. Moreover, emotion detection is'nt limited to recognizing fundamental psychological variables (happy, sad, or furious); depending on the emotion model, it can go up to a 6-scale or else 8-scale. This emotion word count has computed the most often used emotional terms based on the number of times they have been used.



Website: ijetms.in Issue:1, Volume No.4, January-2020 DOI:10.46647/ijetms.2020.v04i01.003



Fig 7: Emotion word cloud of amazon product alexa

4.4 Hierarchical Dendrograms

Dendrogram is graphic representation of hierarchical connection amid items. The most recurrently produced as outcome of hierarchical clustering. Dendrogram's chief drive is figure out optimal method to allot items to clusters[20]. Words have been assigned to hierarchical dendograms using the Euclidean approach in below fig it has shown hierarchical dendogram of amazon product

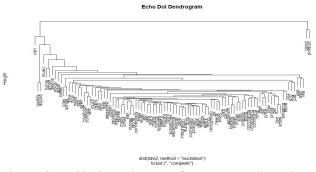


Fig 7: Hierarchical Dendogram of Amazon Product Alexa

4.5 Circular Dendogram

To construct a circular dendogram, also known as a circular, radial, or fan tree plot, use the circlize dendogram function and supply an object to class dendogram. The label track height and dend track height parameters are needed to appropriately fit the labels, and both circlize and dendextend must be installed in order to run the function. All of Alexa's emotive words were checked in a circular dendogram in the below fig it has given in visual that circular dendogram of sentiment analysis of amazon product alexa.

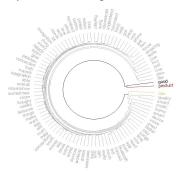


Fig 8: circular Dendogram of Amazon product Alexa

4.6 Word Cloud

Word clouds, also known as tag clouds, are a form of text data visualisation approach that uses the visual depiction of words to express an idea about a subject. They are useful for analysis, surveys, and the collecting of written views, among other things. The word cloud is the most advantageous and easy approach for displaying text data in a graphical manner; not only does a word cloud convey an image of certain words, but it also symbolises something other than the word. The word "cloud" has been used to refer to the first step of a text document's processing in a wide sense. As a result, when classifying or displaying a word cloud, the frequency of terms is the most important factor to consider. After importing the texts, this study employed R packages to pre-process the data, notably for word clouds and frequency visualization[21]. Following the removal of numerals, punctuation, symbols, and hyphens, a lower-case conversion of tokens was conducted using the stemmer in this word cloud of alexa amazon product evaluations.



Fig 9: Word Cloud of Amazon Product Alexa

4.7 Polarity

Each product review that we utilised in Amazon customer reviews was evaluated. We used almost 60 different types of items in this sample. Each product's ratio differs from that of the others. For assessing the amazon product semantical analysis, we looked at polarity. As we all know, there are two concepts that may be used to characterise sentiment analysis: "polarity" and "subjectivity." Individual opinions, thoughts, and feelings are referred to as subjectivity, whereas polarisation simply refers to emotions that are expressed favourably, negatively, or neutrally. Sentiment Classification will address the phrase's scope, the document's scope, and sub-work. We looked at polarity[22] in our investigation.



Website: ijetms.in Issue:1, Volume No.4, January-2020 DOI:10.46647/ijetms.2020.v04i01.003

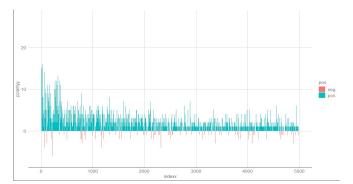


Fig 10: polarity of amazon product alexa

In below fig 11 it has showed the reviews of the amazon product alexa with the polarity and sentiment its negative



Fig 11: reviews of the amazon product alexa with the polarity

4.8 Emotion Mining

In sentiment analysis, studies focused at identifying text by polarity ([positive/Negative]) are the most common. However, a more particular approach based on emotions might be used. This method is still in its infancy in the scientific community (plaza-del-acro 2018). The first issue in such a study is determining what emotions are present, which will define the categorization classes. Two of the numerous emotional models created by psychologists are widely used in emotion identification by computers. Anger, disgust, fear, pleasure, sorrow, and surprise are the six main emotions that Eckman differentiates in his initial model. The other is Plotchik's emotional wheel, which has 8 chief emotions: anger, anticipation, joy, trust, fear, surprise, sadness, as well disgust[23]. In the diagram below, it shows a survey of emotions such as anger, anticipation, disgust, fear, joy, sadness, surprise, as well trust. There will be a count of feelings on various sentiments, and the most common sentiment is trust. The word count of sentiment emotions that has been counted in the sentiment emotions word cloud is shown in the diagram below.

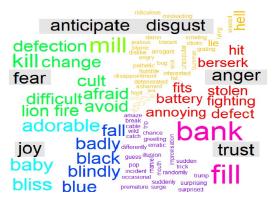


Fig 12:Emotion mining of Amazon Alexa Product

II. RESULTS AND DISCUSSIONS

We can see the reviews by looking at the rating distribution. The majority of reviews were from Amazon's Alexa product, followed by reviews from the Amazon website. Shopping is the second most popular category. We can see from the customer reviews that virtually all of them have a rating of 5,4. Other customers will purchase the goods based on the maximum number of user reviews. It is possible to view the majority of negative and positive terms from evaluations. Exploratory data analysis was effectively shown, and positive and negative terms were discovered. The results of Alexa Product have been exhibited in every approach (Word association, Term Frequency, Word Cloud, hierarchical dendogram, circular dendogram, Polarity, Emotion Mining).

CONCLUSION

In any professional field, data analysis is a critical stage in addressing a client's diverse demands. The diverse range of insights that may be extracted from data is inherently useful, since many firms are actively seeking for futuristic, predictive, and descriptive insights from the raw data that they have previously collected. After the analysis has been completed, it assists companies in gaining access to countless hidden patterns, facts, and pieces of knowledge. We have just completed an analysis utilising Amazon Alexa data that not only gives us with incentives to make wise and intelligent business decisions, but also contributes to the company's general growth. These insights provide diverse stakeholders with a clear view and perspective, allowing them to focus on a good future vision. As long as organisations demand Data Science in their day-to-day decision-making processes, the future scope of Data Analysis will be unaffected. There is also a wide range of options when it comes to generating unique interactive solutions and approaches that are limited to making data exploration considerably more interesting in nature. These continuous developments have solidified a promising future for data analysis as a systemic study that will remain as long as there is a data crunch in any viable field of research in the actual world. Results of each techniques has been given below of every techniques.

REFERENCES

- [1] Roger Peng Author, Exploratory data analysis(2012)
- [2] Benjamin Bengfort author, Applied Text Analysis with Python: Enabling Language Aware Data Products with Machine Learning (2017)
- [3] Allen B. Downey Author, Think Stats: Exploratory Data Analysis by (2014)
- [4] Asghar, Nabiha. (2016). Yelp Dataset Challenge: Review Rating Prediction.
- [5] I. J. Good, "The Philosophy of Exploratory Data Analysis," Philosophy of Science 50, no. 2 (Jun., 1983): 283-295
- [6] Rubayyi Alghamdi and Khalid Alfalqi, "A Survey of Topic Modeling in Text Mining" International Journal of Applications (IJACSA), 6(1), 2015. http://dx.doi.org/10.14569/IJACSA.2015.060121



Website: ijetms.in Issue:1, Volume No.4, January-2020 DOI:10.46647/ijetms.2020.v04i01.003

- [7] Shao, Minglai& Qin, Liangxi. (2014). Text Similarity Computing Based on LDA Topic Model and Word Cooccurrence. 10.2991/sekeie-14.2014.47.
- [8] Arun, R., V. Suresh, C. E. VeniMadhavan and M. NarasimhaMurty. "On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations." PAKDD(2010).
- [9] https://www.yelp.com/dataset/challenge
- [10] Talbot, Justin & Setlur, Vidya & Anand, Anushka. (2014). Four Experiments on the Perception of Bar Charts
- [11] Wang, Li-C. (2016). Experience of Data Analytics in EDA and Test Principles, Promises, and Challenges.
- [12] Heimerl, Florian, Steffen Lohmann, Simon Lange and Thomas Ertl. "Word Cloud Explorer: Text Analytics Based on Word Clouds." 2014 47th Hawaii International Conference on System Sciences (2014): 1833-1842
- [13] S. Lohmann, F. Heimerl, F. Bopp, M. Burch and T. Ertl, "Concentri Cloud: Word Cloud Visualization for Multiple Text Documents," 2015 19th International Conference on Information Visualisation, Barcelona, 2015, pp. 114-120. doi: 10.1109/iV.2015.30.
- [14] Dave, K.; Lawrence, S. & Pennock, D. M. (2003), Mining the peanut gallery: opinion extraction and semantic classification of product reviews., in 'WWW', pp. 519-528.
- [15] Pang, Bo & Lee, Lillian & Vaithyanathan, Shivakumar. (2002). Thumbs up? Sentiment Classification Using Machine Learning Techniques. EMNLP. 10
- [16] R programming greeks of greeks
- [17] Sentiment Analysis of Movies on Social Media using R Studio, International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249-8958 (Online), Volume-8 Issue-6
- [18] Text mining and sentiment Analysis :Analysis with R, reggate hub
- [19] Semantic analysis of big data with an emphasis on sentiment analysis from wed based data
- [20] display R Dendogram
- [21] Sentiment Analysis of Amazon Products Using Ensemble Machine Learning Algorithm, International Journal of Mathematical, Engineering and Management Sciences Vol. 4, No. 2, 508–520, 2019 https://dx.doi.org/10.33889/IJMEMS.2019.4.2-041
- [22]] Amazon Product Sentiment Analysis using Machine Learning Techniques
- [23] Sentiment Analysis of Text Documents, Proceedings of the Central European Conference on Information and Intelligent Systems.