|  |  |
| --- | --- |
| **2022- 20223** | **MSc Dissertation** |

|  |  |
| --- | --- |
| **Dissertation Title** | **Unveiling Insights from Amazon Reviews: A Multidimensional Analysis with Sentiment Assessment, and Topic modelling** |

|  |  |
| --- | --- |
| **Author** | **Ben George** |
| **K-Number** | **K00278093** |

**MSc** - **Data Analytics (Level 9)**

|  |  |
| --- | --- |
| **Award Title** | **MSc in Data Analytics** |
| **Submitted to:** | **Technological University of the Shannon: Midlands,**  **(27-08-2023)** |

Text

Description automatically generated

## Declaration of Ethical Practice

Definition: (The Technological University of the Shannon: Midlands)

“The deliberate or reckless representation of another’s words, thoughts, or ideas as one’s own without attribution in connection with submission of academic work, whether graded or otherwise”

As an aspiring professional I am aware of my responsibility to maintain a high level of integrity and to promote the standing of my profession.

Arising from this responsibility I wish to affirm that this dissertation is the result of my own effort and that I have rigorously referenced and acknowledged all sources of information, writing and ideas used in this dissertation.

|  |  |
| --- | --- |
| **Author** | **Ben George** |

|  |  |
| --- | --- |
| **Signature** |  |

|  |  |
| --- | --- |
| **Date** | **27-08-2023** |

## Acknowledgements

I would like to sincerely appreciate the continued support and invaluable guidance from my supervisor David Leonard for his relentless guidance, constructive criticism, advice, and patience. I wish to thank TUS in general for granting me the opportunity to pursue this graduate study.

To my family members for the support and encouragement accorded to me, words cannot express the gratitude I have for you.

Finally, to my classmates and colleagues who encouraged me when I felt I was not equal to the task, thank you so much and may God bless you all.

## Table of Contents

[Declaration of Ethical Practice i](#_Toc143992712)

[Acknowledgements ii](#_Toc143992713)

[Table of Contents iii](#_Toc143992714)

[List of Figures v](#_Toc143992715)

[Glossary of Abbreviations vii](#_Toc143992716)

[Abstract viii](#_Toc143992717)

[Chapter 1 | Introduction 10](#_Toc143992718)

[1.1 Topic Introduction 10](#_Toc143992719)

[1.2 Motivation 11](#_Toc143992720)

[1.3 Research aims and objectives 11](#_Toc143992721)

[1.4 Research questions 12](#_Toc143992722)

[1.5 Thesis structure 12](#_Toc143992723)

[Chapter 2 | Critical Review of Literature 14](#_Toc143992724)

[2.1 Introduction 14](#_Toc143992725)

[2.2 Influence of Online Reviews on Consumer Purchase Behaviour 14](#_Toc143992726)

[2.3 How negative online reviews change customer perception 15](#_Toc143992727)

[2.4 Data collection and data preprocessing 15](#_Toc143992728)

[2.4.1 Data collection methods 16](#_Toc143992729)

[Python Libraries: 16](#_Toc143992730)

[Third-Party Web Scraping Tools: 16](#_Toc143992731)

[Amazon Product Advertising API: 16](#_Toc143992732)

[Selenium Web Scraping: 17](#_Toc143992733)

[2.5 Data preprocessing 17](#_Toc143992734)

[Lower text: 17](#_Toc143992735)

[Removal of @mention: 17](#_Toc143992736)

[Removal of URL: 17](#_Toc143992737)

[Punctuation Removal: 18](#_Toc143992738)

[Removing whitespace: 18](#_Toc143992739)

[Removing Numbers: 18](#_Toc143992740)

[Tokenizing: 18](#_Toc143992741)

[Removing Stop words: 18](#_Toc143992742)

[Stemming: 18](#_Toc143992743)

[Count vectorizer: 18](#_Toc143992744)

[2.6 Sentiment analysis 19](#_Toc143992745)

[2.6.1 Different Sentiment analysis approaches 19](#_Toc143992746)

[Lexicon-based (Unsupervised) Approach: 20](#_Toc143992747)

[Machine Learning (Supervised) Approach: 20](#_Toc143992748)

[Deep Learning-Based Approach: 20](#_Toc143992749)

[Rule based approach: 20](#_Toc143992750)

[Hybrid approach: 21](#_Toc143992751)

[2.6.2 Valence Aware Dictionary and sentiment Reasoner (Vader) 21](#_Toc143992752)

[2.6.3 Word order sensitivity rules of Vader model 22](#_Toc143992753)

[2.7 Topic modelling 23](#_Toc143992754)

[2.7.1 Topic modelling terms 23](#_Toc143992755)

[2.7.2 Topic modelling Latent Dirichlet Allocation approach (LDA) 23](#_Toc143992756)

[2.8 Related works 24](#_Toc143992757)

[2.8.1 Sentiment analysis in social media platforms 24](#_Toc143992758)

[2.8.2 Sentiment analysis approach using POS tagging 25](#_Toc143992759)

[2.8.3 Sentiment analysis approach using Support Vector Machine Classifier (SVM) 26](#_Toc143992760)

[2.8.4 Sentiment analysis approach using Latent Dirichlet Allocation 27](#_Toc143992761)

[2.9 Research gaps 27](#_Toc143992762)

[2.10 Conclusion 28](#_Toc143992763)

[Chapter 3 | Research Methodology 30](#_Toc143992764)

[3.1 Introduction 30](#_Toc143992765)

[3.2 Business understanding 30](#_Toc143992766)

[3.3 Proposed methodology 31](#_Toc143992767)

[3.4 Data collection 32](#_Toc143992768)

[3.4.1 Data selection criteria 33](#_Toc143992769)

[3.4.2 Dataset description 35](#_Toc143992770)

[3.5 Data Preprocessing 37](#_Toc143992771)

[3.5.1 Basic data preprocessing steps. 38](#_Toc143992772)

[3.5.2 Deeper data preprocessing steps. 39](#_Toc143992773)

[3.6 Exploratory Data analysis (EDA) 40](#_Toc143992774)

[3.7 Classification of reviews into product and service reviews 47](#_Toc143992775)

[3.8 Sentiment analysis modelling using Vader 49](#_Toc143992776)

[3.9 Topic modelling 51](#_Toc143992777)

[3.9.1 Identify the optimal number of topics 52](#_Toc143992778)

[3.9.2 visualise the LDA model using pyLDAvis visualisation library 52](#_Toc143992779)

[3.10 Summary 53](#_Toc143992780)

[Chapter 4 | Research Findings 55](#_Toc143992781)

[4.1 Introduction 55](#_Toc143992782)

[4.2 Classification of amazon reviews to product and Service reviews 55](#_Toc143992783)

[4.3 Sentiment analysis using Vader algorithm 57](#_Toc143992784)

[4.4 Introducing a parallel sentiment rating for amazon 64](#_Toc143992785)

[4.4.1 Analysis for Xiaomi 11 Lite NE 5G (Jazz Blue 6GB RAM 128 GB Storage) 64](#_Toc143992786)

[4.4.2 Analysis for Samsung Galaxy M33 5G (Deep Ocean Blue, 8GB, 128GB Storage) 66](#_Toc143992787)

[4.4.3 Analysis for OnePlus 9R 5G (Carbon Black, 12GB RAM, 256 GB Storage) 67](#_Toc143992788)

[4.5 Topic modelling 69](#_Toc143992789)

[4.6 Tuning the topic model (LDA) to achieve maximum accuracy 70](#_Toc143992790)

[4.7 Analysing Topic Bubble Chart 71](#_Toc143992791)

[4.8 Analysing pyLDAvis visualization graph 73](#_Toc143992792)

[4.9 Summary 75](#_Toc143992793)

[Chapter 5 | Conclusions and Recommendations 77](#_Toc143992794)

[5.1 Conclusion 77](#_Toc143992795)

[5.2 Answers derived to research questions 77](#_Toc143992796)

[5.3 Limitations 79](#_Toc143992797)

[5.4 Future work 79](#_Toc143992798)

[Appendix 1 80](#_Toc143992799)

[References 83](#_Toc143992800)

## List of Figures

[Figure 1,Different sentiment analysis approaches (Kausar et al,2019) 19](#_Toc143989718)

[Figure 2, Proposed methodology flow chart 31](#_Toc143989719)

[Figure 3,Data collection procedure 34](#_Toc143989720)

[Figure 4,Data extraction procedure, detailed flowchart 35](#_Toc143989721)

[Figure 5,Dataset description 36](#_Toc143989722)

[Figure 6,Dataset sample 36](#_Toc143989723)

[Figure 7,Unigram graph based on amazon reviews 41](#_Toc143989724)

[Figure 8,bigram graph based on amazon reviews 41](#_Toc143989725)

[Figure 9,Trigram graph based on amazon reviews 42](#_Toc143989726)

[Figure 10, Character count, word count, word density, punctuation count of amazon reviews 43](#_Toc143989727)

[Figure 11, Overall review rating for each company 44](#_Toc143989728)

[Figure 12, Total ratings given for each model 44](#_Toc143989729)

[Figure 13, Average number of helpful ratings for each model 45](#_Toc143989730)

[Figure 14, Total number of emoticons in reviews for each model 46](#_Toc143989731)

[Figure 15, Word cloud for Overall reviews 46](#_Toc143989732)

[Figure 16,Product review and service review labelled data 56](#_Toc143989733)

[Figure 17,Overall product and service review 56](#_Toc143989734)

[Figure 18, Product and service reviews for each model 57](#_Toc143989735)

[Figure 19,Dataset with product review and service review labels 58](#_Toc143989736)

[Figure 20,distribution of sentiments for overall reviews in the dataset 59](#_Toc143989737)

[Figure 21,sentiment score distribution for all the reviews in the dataset 60](#_Toc143989738)

[Figure 22,Averege sentiment score against the review rating 61](#_Toc143989739)

[Figure 23,Sentiment proportion for each smartphone model in the dataset 62](#_Toc143989740)

[Figure 24,Sentiment distribution based on the product reviews and service reviews available in the review text from the whole dataset 62](#_Toc143989741)

[Figure 25, Corelation matrix Heatmap 64](#_Toc143989742)

[Figure 26,Total product and service reviews for Xiaomi 11 Lite NE 5G 65](#_Toc143989743)

[Figure 27,Product review and service review sentiment analysis for the Xiaomi 11 Lite NE 5G 65](#_Toc143989744)

[Figure 28,Most frequently discussed words for the Xiaomi 11 Lite NE 5G 66](#_Toc143989745)

[Figure 29,Total product and service reviews for Samsung Galaxy M33 5G 66](#_Toc143989746)

[Figure 30,Product review and service review sentiment analysis for the Samsung Galaxy M33 5G 67](#_Toc143989747)

[Figure 31,Most frequently discussed words for the Samsung Galaxy M33 5G 67](#_Toc143989748)

[Figure 32,Total product and service reviews for OnePlus 9R 5G 68](#_Toc143989749)

[Figure 33,Product review and service review sentiment analysis for the OnePlus 9R 5G 68](#_Toc143989750)

[Figure 34,Most frequently discussed words for the OnePlus 9R 5G 69](#_Toc143989751)

[Figure 35,Line graph with coherence score for 5 topics 70](#_Toc143989752)

[Figure 36,LDA Bubble chart 72](#_Toc143989753)

[Figure 37,LDA most discussed topics in Topic and Topic 2 73](#_Toc143989754)

[Figure 38, LDA most discussed topics in Topic 3 and Topic 4 74](#_Toc143989755)

## Glossary of Abbreviations

|  |  |
| --- | --- |
| **Abbreviation** | **Full Title** |
| NLP | Natural Language Processing |
| VADER | Valence Aware Dictionary and Sentiment Reasoner |
| eWOM | Electronic word of mouth |
| JSON | JavaScript Object Notation |
| SVM | Support Vector Machines |
| RNN | Recurrent Neural Networks |
| POS | Parts of speech |
| TF-IDF | Term Frequency-Inverse Document Frequency |
| SGD | Stochastic Gradient Descent |
| TDS | Topic/Document/Sentence |
| JST | Joint sentiment topic |
| EDA | Exploratory data analysis |
| NLTK | Natural Language Toolkit |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

## 

## Abstract

In the digital era most of the purchase is happening thought online platforms this digital landscape of ecommerce, Amazon stands a prominent online platform where consumers reviews play an important role in impacting purchase behaviours. These reviews evaluate both the product features and the quality of the service provided by amazon. To provide more informed decision making, this paper presents a combined approach that involves categorizing customer reviews into product reviews and service reviews. The proposed system employs Valence Aware Dictionary and sentiment Reasoner (VADER) algorithm to determine sentiment polarity, and it leverages Latent Dirichlet Allocation (LDA) for topic modelling. This strategy gives customers an understandable distinction between product and service evaluations, improving their capacity to judge which reviews are relevant to their requirements. Users are able to comprehend general sentiment patterns related to products and services by using the sentiment analysis, which measures the emotional tone of reviews. By identifying underlying themes in reviews, the LDA-based topic modelling enhances the analysis even further. The result of the combined system is a solid solution that enables customers to make knowledgeable decisions as they explore the dynamic world of Amazon reviews. The online shopping experience is improved by the combination of sentiment analysis and topic modelling, which leads to a more detailed knowledge of customer opinions and interests.

|  |  |
| --- | --- |
| **Chapter 1** | **Introduction** |

# | Introduction

## 1.1 Topic Introduction

Customer reviews are becoming an important pillar for companies in the digital age. They offer important input that can influence product development, marketing plans, and customer service enhancements. Parsing through each review manually would be unfeasible given the enormous volume of Amazon reviews that are generated every day. Here is where sentiment analysis, an effective method for understanding and employing customer feedback, comes into action.

Sentiment analysis, commonly referred to as opinion mining, is a branch of Natural Language Processing (NLP) that focuses on locating and obtaining subjective data from sources. Sentiment analysis makes it possible for computers to understand and analyse human emotions and views by examining the sentiment expressed in text data. Sentiment analysis classifies feedback in the context of Amazon reviews as positive, negative, or neutral, giving organizations useful information about customer sentiment. The VADER (Valence Aware Dictionary and Sentiment Reasoner) algorithm is one of the most widely used sentiment analysis methods. A vocabulary of words and phrases is used by the rule-based method VADER to analyse a text's sentiment. The analysis of social media data and product reviews, including those on Amazon, has shown that this algorithm is reliable. Numerous studies using various NLP methods including the VADER algorithm have been done on sentiment analysis of Amazon reviews. (Abah,2021) As an illustration of the great accuracy that deep learning models may attain in sentiment analysis, Abah used deep learning methods to evaluate Amazon electrical product evaluations. (Rashed,2021) Another study by Rashed Hamdallah looked at how sentiment analysis using Amazon reviews can offer insightful information about customer sentiment, helping businesses better understand the preferences and opinions of their customers.

## 1.2 Motivation

It is crucial for businesses to categorize Amazon reviews into product and service reviews since doing so enables them to learn more about how customers feel about particular features of their services. Businesses can determine their strengths and shortcomings in each area by examining the sentiment of product ratings independently of service feedback. Following that, this information may be used to allocate resources wisely, raise consumer happiness, and modify marketing plans as necessary. The most talked-about phrases in Amazon reviews can also be used as potent selling points or advantages to draw in buyers. Companies can take advantage of features when a product receives several positive reviews that highlight those qualities in their marketing campaigns. Sentiment research is essential in finding these characteristics so that firms can emphasize them and profit from their appeal to prospective customers.

## 1.3 Research aims and objectives

The purpose of the study is to examine the variables affecting consumer loyalty in online buying as well as the elements affecting customer happiness. The purpose of the study is to better understand the aspects that influence customers' happiness with their online shopping experiences and how these elements affect their inclination to stick with a certain brand or retailer. This study aims to shed light on the relationships between customer satisfaction and loyalty in the context of online shopping, helping to better understand consumer behaviour in the virtual economy.

The research objectives incudes:

1. To classify the reviews into product reviews and service reviews to identify the major discussion in both areas separately.
2. Previous research has demonstrated sentiment analysis for Twitter data by determining the opinions of customers toward certain products. This work gathers and analyse the reviews from amazon to determine the sentiment using lexicon-based VADER algorithm.
3. To use topic modelling techniques like Latent Dirichlet Allocation (LDA), helps identify underlying themes or topics within a collection of documents. This is quite helpful for categorizing information, spotting trends, and comprehending the key topics covered in a huge corpus of text.

## 1.4 Research questions

Research question are formulated to guide identify scope of our study and provide a roadmap for the research process. Following research questions, a created to fulfil this purpose:

* 1. How previous customer reviews can affect the future customer purchase behaviours?
  2. How does sentiment differ between Amazon product and service reviews, and what insights can be gained from this analysis?
  3. How effective is the VADER algorithm in analysing sentiment in Amazon product and service reviews?
  4. How does topic modelling help in identifying hidden semantic structures in Amazon reviews?
  5. How can sentiment analysis and topic modelling be used to improve targeted advertising on Amazon?

## 1.5 Thesis structure

An explanation of this thesis's structure is provided below. Chapter 2 presents the literature study on Amazon reviews and various methods to conduct sentiment analysis. Chapter 3 discusses dataset extraction, sentiment extraction techniques, and topic modelling techniques. The findings and analysis from the results are presented in chapter 4. Chapter 5 presents conclusions and limitations.

|  |  |
| --- | --- |
| **Chapter 2** | **Critical Review of Literature** |

# | Critical Review of Literature

## 2.1 Introduction

In the current digital era, a significant portion of consumers worldwide engage in online purchasing. By the end of 2023, Forbes Advisor projects that global e-commerce revenues will amount $6.310 trillion. (Anna Baluch, 2023). The most popular online retailer among these e-commerce sites is Amazon, which sends out about 1.6 million shipments every day. Amazon has millions of product options and user base which can significantly influences consumer purchase decisions. Customer reviews and ratings on Amazon contain useful information regarding the quality of product and services offered by Amazon. These customer testimonials are sources that can reveal details about the sentiments, opinions and experiences of consumers.

Sentiment analysis is a field of natural language processing (NLP), which is also known as opinion mining. This method extracts polarity and subjectivity of information from the given text. In the context of amazon reviews, sentiment analysis is a powerful tool which can be used by both sellers and consumers. For sellers, the results can be used to improve customer services and provide more customer satisfaction. On the other hand, for consumers, it aids in making an informed decision in purchasing a product.

This literature review aims to explore the existing research and methodologies regarding the sentiment analysis on Amazon product reviews. By analysing the finding of previous studies, this review seeks to shed light some on to the different techniques used, challenges encountered and the advancements made in this field.

## 2.2 Influence of Online Reviews on Consumer Purchase Behaviour

The development of e-commerce websites is leading customers to do more online shopping than ever. Customer reviews are a major factor that can affect sales of a product in a positive or negative way. Reviews can be considered as an evaluation information of the products and services based on customer experiences. Customer tend to infer the quality of the product based on the ratings and reviews, which will help them to save personal time and risk of purchase (Mo, Li and Fan 2015). The expansion of internet in recent years helped the sellers in using electronic word of mouth (eWOM) as a way to advertise the products. Although the eWOM is used widely, customers tend to rely more on reviews to take a final purchase decision (Chevalier and Mayzlin, 2006). Online reviews help the customers to understand the quality of the products from a customer perspective (Bickart and Schindler, 2001). Recent studies and surveys show that 83% of the customers purchase decisions are entirely dependent upon the product reviews (Zhang et al., 2014).

## 2.3 How negative online reviews change customer perception

The studies performed by Tang clearly shows that the negative reviews about a product can create a negative perception towards the product (Tang et al., 2014). Negative reviews have a greater impact than positive reviews, this behaviour is more related to human psychology. Many researches were conducted to unravel the mystery behind the impact on customer perception by negative product reviews. Customers are more drawn to negative reviews, because it contains more information for an effective decision making (Maheswaran and Meyers-Levy, 1990).

(Kahneman and Tversky, 1979) Kahneman and Tversky they formed a psychological concept called Prospect Theory, that explains how individuals make decisions under conditions of risk and uncertainty. According to the hypothesis, people consider possible losses more seriously than benefits of a same magnitude. A S-shaped curve with higher slopes for losses than for profits serves as the prospect theory's representation of the value function. In this context, the above theory can be interpreted as follows. The negative comments by the previous customers can convey a negative sentiment to the product or service to the future customers. These negative perceptions are weighed more than the gains of positive reviews.

## 2.4 Data collection and data preprocessing

It is clear from the studies mentioned above how customer purchase behaviour is influenced by product reviews. Understanding the emotions behind previous customer interactions is crucial for consumers and customers. The sentiments hidden in a text can be gleaned using a variety of techniques. We will talk about these approaches in greater detail in this section, which will make it clearer to us which approach is best for our cause.

### 2.4.1 Data collection methods

### Python Libraries:

Python modules like Beautiful Soup and Requests can be used to scrape product reviews from Amazon.com. While Requests is used to send HTTP requests, Beautiful Soup is used to parse HTML and XML documents. An account need be created on these websites to use their APIs. Request and Beautiful Soup are two different tools in Python that help gather all the organized and messy information from a product page. They collect codes and webpage parts. At the same time, Beautiful Soup looks at this collected stuff and gets all the details from it, like Amazon reviews using a special number called ASIN - Amazon Standard Identification Number. Then, the data changes into a format like CSV, JSON, or XLS, gets saved, and you can use it later. (Nguyen et al,2018).

### Third-Party Web Scraping Tools:

To extract data from websites, like Amazon, there are third-party web scraping tools and platforms that provide point-and-click user interfaces. These programs can save you time setting up the scraping process and frequently require less programming skills. One of the popular third-party web scraping tools available is Web Scraper (2020). Web Scraper (2020), a third-party cloud-based web scraping service, due to its practical low-code nature and our existing expertise with its desktop-product. Definitions for scrape jobs can be created in the Chrome browser and exported in JSON format. Scrape tasks can navigate paginated web sites, drill through pages, and collect relevant data elements like product, review, and reviewer. There are features for robot identification and avoidance, such as pauses between page requests and the usage of different IP addresses. A complete API is offered for programmatic job definition in the cloud-based service (Woodall et al,2021).

### Amazon Product Advertising API:

Access to a variety of Amazon product data, including reviews, is made possible by the Product Advertising API of Amazon. However, usage restrictions and constraints may apply to API access. The option to download Amazon reviews is available in JSON format. Because JSON (JavaScript Object Notation) relies on a key-value pairing system, which starts in an unstructured format, the information is adapted into a CSV file to enhance the effectiveness of preprocessing procedures.

### Selenium Web Scraping:

Selenium is a well-liked web scraping tool that enables automated website interaction. By identifying and parsing HTML components, Selenium may be used to navigate to Amazon product pages, scroll through reviews, and retrieve review data. This method offers extensive control over the extraction process but may necessitate additional programming expertise. The most practical usage of Selenium is when dealing with interactive, JavaScript-heavy pages, such as those on travel websites like TripAdvisor, Airbnb, and Expedia. You can scrape anything that is shown on the screen using other Python tools (such as BeautifulSoup, request, Scrapy, etc.). Information that is hidden cannot be scraped until you click on it. The HTML tags of a website can be used to extract its data. Use of the XPATH allows access to the precise location (Han, S. and Anderson, C.K., 2021).

## 2.5 Data preprocessing

An important part of sentiment analysis is data preparation, which entails converting unstructured data into an analytically useful format. Data integration, data reduction, data cleaning, and data transformation are just a few of the steps that make up the data preprocessing process. Remove redundant or irrelevant data, fix errors, and deal with missing data are all parts of data cleaning. Creating a single dataset through data integration involves merging data from several sources. Data transformation involves transforming data into a form that is suitable for analysis, such as translating text data into numerical data. Reducing the dataset's size while preserving crucial information is known as data reduction.

### Lower text:

Convert all the words to lowercase format to ensure uniformity and reduce the vocabulary size. It assists in lowering the number of words the dictionary must hold at one time (Pradha, Halgamuge and Vinh,2019).

### Removal of @mention:

It makes it easier to remove user mentions because they don't add any valuable context to the text's sentiment (Pradha, Halgamuge and Vinh,2019).

### Removal of URL:

Any URLs should be removed from the text. It involves removing URLs with the HTTP, https, and image prefixes (Pradha, Halgamuge and Vinh,2019).

### ****Punctuation Removal:****

This is a process of removing any punctuation or alphabetic words that don't have a context in the sentiment of the original text (Pradha,Halgamuge and Vinh,2019). These unnecessary symbols create noise in text modeling. To handle this, you can remove these symbols using Python's built-in 're' function, which utilizes regular expressions.

### Removing whitespace:

Whitespace is deleted for computational reasons because it adds no value to the text (Pradha,Halgamuge and Vinh,2019).

### Removing Numbers:

Text has numbers, but numbers don't have sentiments. Deleting numbers from the data during preprocessing is crucial (Thakkar et al,2022).

### Tokenizing:

Tokenizing is thought to be a crucial pre-processing step in the classification of text data. It is quite difficult to train an algorithm to categorize text data using a whole page or a single sentence. Tokenizing the phrase into words is therefore required in order to train the classifier on positive and negative terms (Chaithra, V.D., 2019).

### Removing Stop words:

These are the words that don't really contribute much to whether we label anything as a positive or bad word. Stop words in English (a, as, is, the, etc.) are eliminated from the sentences because they are not required for categorizing the sentence or document (Chaithra, V.D., 2019).

### Stemming:

Finding a word's root is a procedure called stemming. The Porter Stemmer algorithm is used for this. This cuts down on the algorithm's training time for both the positive and negative tenses of the term (Chaithra, V.D., 2019).

### Count vectorizer:

Words are converted into vectors that a machine learning system can easily grasp with the aid of a count vectorizer. When count vectorizer is applied to the data, a vector matrix for each word in the dataset is produced (Chaithra, V.D., 2019).

## 2.6 Sentiment analysis

In Natural Language Processing (NLP), sentiment analysis is a field that involves computerised analysis the emotions behind a text (M. Bautin, L. Vijayarenu and S. Skiena,2008). sentiment analysis has become an essential instrument for organizations, governments, and researchers to comprehend the general public's views, client feedback, and social trends. Sentiment analysis has become an essential instrument for organizations, governments, and researchers to comprehend the general public's views, client feedback, and social trends. In now days sentiment analysis is a popular tool to analyse the election results (Bollen, J., Mao, H. and Zeng, X., 2011) or predicting the stock market based on social media comments (Xu, T., Peng, Q. and Cheng, Y., 2012). It involves computational methods to identify the sentiment behind a text to classify as positive, neutral or negative.

### 2.6.1 Different Sentiment analysis approaches

There are several ways to identify the sentiment behind a text. Based on the dataset the approach may differ. Below are the common sentiment analysis approaches.

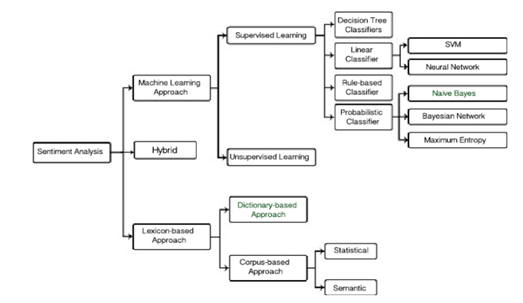


Figure 1,Different sentiment analysis approaches (Kausar et al,2019)

### Lexicon-based (Unsupervised) Approach:

(Ding, X., Liu, B. and Yu, P.S., 2008) In this study, Ding, Liu, and Yu describe a comprehensive lexicon-based approach for sentiment analysis, also known as opinion mining. The method makes use of a sentiment lexicon, a dictionary that lists words together with the associated sentiment scores. By conducting a document-level analysis of the text with the goal of understanding the overall sentiment conveyed in the entire text. They determine whether a sentiment is positive, negative, or neutral by adding up the sentiment ratings assigned to each word in the text. The drawback of this method is, it cannot handle the context and expression of the given text.

### Machine Learning (Supervised) Approach:

Based on training data that has been tagged, machine learning algorithms can be taught to categorize the sentiment of a statement. With this strategy, different linguistic properties including n-grams, POS tags, and grammatical dependencies are extracted from the text. After that, these features are fed into classifiers like Naive Bayes, Support Vector Machines (SVM), or Recurrent Neural Networks (RNN) in order to train them to predict sentiment (Copaceanu, 2021; Alharbi et al,2021).

### Deep Learning-Based Approach:

To determine the tone of a text, deep learning models like Recurrent Neural Networks (RNNs) and Transformer models are used. These models are able to determine the context and complex connections that exist within a sentence. Deep learning-based sentiment analysis can be seen as a classification task. This method trains the model to divide the text into categories that are either positive negative, or neutral.

### Rule based approach:

This approach involves in creating certain handcrafted rules. These rules are based on the grammatical structure or specific words associated with sentiment. For example, a negotiation word like “not” or” no” can be considered as a negative sentiment. The approach involves various preprocessing methods such as part-of-speech tagging, adverbs and adjectives collection. Once the preprocessing is completed a scoring method is used to classify the method into a seven-point scale (Almost Positive, Positive, Very Positive, Almost Negative, Negative, Very Negative and Neutral). Senti-Wordnet tool can be used to identify the score and polarity of each word (Ray, P. and Chakrabarti, A., 2022).

### Hybrid approach:

(Chaithra, V.D, 2019) In this paper V.D. Chaithra describes the method to combine the Naive Bayes algorithm and the lexicon-based Vader algorithm to identify the sentiment of a text. Naive Bayes calculates the probability of a text being positive, negative, or neutral sentiment) based on the occurrence of words or features in the text. The combination of the Naive Bayes classifier and the VADER sentiment analyser increases the effectiveness of the sentiment analysis.

These days, numerous independent libraries are available with the above-mentioned techniques combined. These libraries are used in different programming languages based on the analysis approaches and nature of the dataset. Some of the popular libraries are TextBlob, VADER, AFINN and Sentiment intensity analyser.

### 2.6.2 Valence Aware Dictionary and sentiment Reasoner (Vader)

VADER is a sentiment analysis tool that uses rules-based approach and lexicon approach combined developed by Hutto and Gilbert (Hutto C and Gilbert E, 2014). The common format of social media content is a major obstacle in analysing its sentiment. Hence Hutto and Gilbert introduced the Vader, a simple model that can be used for sentiment analysis. Vader is evaluated against benchmarking standards such as SentiWordNet, ANEW, LIWC, Support Vector Machine and Entropy. Vader was able to outperform the Human raters with a F1 classification accuracy of 0.96 and 0.84.

Vader is useful to handle terms, acronyms, slang, emoticons, and emojis often used in social media. When compared to machine learning techniques, rule-based sentiment analysis is popular for its effectiveness and quickness. Rule-based sentiment analysis depends on established rules and linguistic patterns as opposed to machine learning approaches, which need training on big datasets to identify patterns and features. A vector of sentiment scores with negative, neutral, positive, and compound polarities label is generated for each text. The normalised values of polarities range from 0 to 1 and the compound score varies in between −1 (negative) to 1 (positive) (Hutto C and Gilbert E, 2014). The Vader Python libraries are free software under the MIT license. Vader is now freely available on the marketplace for anyone to use as a result.

Hutto and Gilbert's key arguments when Vader is presented were as follows:

*“1) works well on social media style text, yet readily generalizes to multiple domains,*

*2) requires no training data, but is constructed from a generalizable, valence-based, human-curated gold standard sentiment lexicon*

*3) is fast enough to be used online with streaming data*

*4) does not severely suffer from a speed-performance trade-off.” (C.J. Hutto,Eric Gilbert, 2014).*

### 2.6.3 Word order sensitivity rules of Vader model

(*C.J. Hutto,Eric Gilbert, 2014)*)The authors used qualitative analysis to identify the change of sentiments of a sentence based on the grammatical structure of English language.

1. Punctuation, without changing the semantic orientation, punctuation, namely the exclamation mark (!), increases the amount of the intensity. For instance, "The food here is good!!!" is more intense statement than "The food here is good."

2. Capitalization, when a sentiment-relevant word is highlighted in capital letters in comparison to other non-capitalized terms, the sentiment intensity increases without changing the meaning of the sentence. As in, "The food here is GREAT!" conveys more emotion than "The food here is delicious!"

3. Degree modifiers also play a major role in intensifying the sentiment of a sentence. They are also known as intensifiers or booster words. When compared to "The service here is good," for instance, "The service here is extremely good" is more intense, but "The service here is marginally good" is slightly less in intensity.

4. Trigrams will help us identify 90% of the cases that are converted to negative sentiment due to the use of negation words. A good example for negated sentence is “The food here isn’t really all that great”.

5. A sentence can have a mixed sentiment fur to the use of contrastive conjunction like “but”. This cause change in polarity of the sentence, due to the important nature of sentence followed by the conjunction

The authors added additional 7500 features to the word bank. This made the Vader model to analyse the social media text more effectively. The Vader model is tested against 400 positive and 400 negative tweets to analyse its accuracy. All these properties make Vader a perfect option for sentiment analysis for Amazon product reviews.

## 2.7 Topic modelling

Natural language processing (NLP) techniques like topic modelling are used to find hidden themes or subjects in a group of documents (Rani and Kumar,2021; Liu et al,2016). In order to group related documents together according to their content, it includes finding patterns in the text data. Latent Dirichlet Allocation (LDA), which extracts topic probabilities from statistical data, is the most widely used topic modelling algorithm. Information retrieval, query expansion, classification, categorization, and document summarizing all use topic modelling Topic modelling does present certain difficulties, such as selecting the number of samples and topics that need to extract (Weston et al,2023). Working with brief text is less beneficial for topic modelling than working with long-format data (Silva et al,2021.

### 2.7.1 Topic modelling terms

There are several terms related topic modelling which need to understand while using the model. Vocabulary is the collection of distinct words in the corpus. Corpus is a set made up of M documents that contains the whole data in the dataset. Term Frequency-Inverse Document Frequency (TF-IDF) is a method used to determine word relevancy in topic modelling (Vayansky and Kumar,2020). Topic is a probability distribution across the vocabulary, representing different themes or topics within the corpus (Blei et al,2003).

Bag-of-words is the representation of documents used in topic modelling, where the order of words is ignored (Arora et al,2013).

### 2.7.2 Topic modelling Latent Dirichlet Allocation approach (LDA)

A three-level hierarchical Bayesian model called Latent Dirichlet Allocation (LDA) is employed in topic modelling Each object in a collection, like a document, is modelled in LDA as a finite mixture over an underlying set of subjects. It is a generative model that depicts documents as a combination of subjects and topics as a combination of words. LDA makes it possible to find hidden themes in a group of documents by estimating the topic and word distributions (Blei,2003).

The study presented by makes a contribution to the subject of tourism management by using LDA to extract useful data from online ratings and reviews, offering insightful data on visitor satisfaction. Based on the characteristics that have been identified as impacting satisfaction, the results can assist tourism businesses and destinations improve their products and enhance visitor experiences (Guo and Barnes,2016). Overall, the paper demonstrates the effectiveness of LDA in analysing online ratings and reviews for understanding tourist satisfaction, offering a valuable approach for extracting insights from large volumes of unstructured text data.

## 2.8 Related works

Sentiment analysis is used in a variety of fields where understanding people's attitudes, feelings, and views is crucial. The several domains that are similar to Amazon product review sentiment analysis and the research done there can be discussed in more detail in this section.

### 2.8.1 Sentiment analysis in social media platforms

(Hasan, A., Moin, S., Karim, A. and Shamshirband, S., 2018) The authors investigate methods and procedures for gathering sentiments and viewpoints from Twitter posts. Twitter is an established social media platform where users post short text messages expressing their feelings, thoughts, and opinions. They were able to gather 100,000 tweets from public twitter accounts using Tweepy API. Data was gathered in order to reveal the opinions of the two largest democratic parties in India. This study compares a variety of sentiment analysers, particularly sentiment lexicons (W-WSD, SentiWordNet, TextBlob), and machine learning algorithms (Nave Bayes and SVM) to identify the strategy that is most accurate in identifying voters' attitudes toward upcoming elections. The authors used two machine-learning classifiers (Naive Bayes and SVM) to assess their findings after validating three sentiment analysis lexicons (SentiWordNet, TextBlob, and W-WSD). After preprocessing the data out of 100,000 tweets only 6250 rows of tweets remained. The cleaned data is then fed to lexicon analysis tools. After the analysis text blob was able to predict 54.08%, SentiWordNet predicted 41.488% and W-WSD predicted 51.024% of the tweets as positive. Naive Bayes classifier and SVM algorithm were able to cross check the accuracy of the result provided by the lexicon tools. Naive Bayes classifier predicted an accuracy of 76% for text blob and 79% for Wordnet respectively. Based on the results, the authors concluded that, Text blob and W-WSD can predict sentiment more accurately.

(Naf'an et al,2019) Instagram users have the option to like, share, and comment on other people's posts as well as their own. Even though it is against the law, the comment section is frequently abused and leads to cyberbullying. Instagram, unfortunately, does not have a feature to identify cyberbullying. The goal of this project is to create a system that can categorize comments and tell whether or not they have aspects of cyberbullying. The data required for the analysis extracted from a personal Instagram account using Instagram API library. The data will be manually labelled using the data that has been gathered using a bag of words that are related to cyber bullying. The author receives a total of 455 comments that have been labelled, including 284 comments that do not contain any cyberbullying-related features and 171 comments that do. Based on statistical measurements that show the frequency of occurrence of a word in the document, TF-IDF will provide word weighting. A word's value of conformance will increase with its level of usage in the comments. If a sentence contains a repeating word from the bag of words, the weightage of overall sentence will increase. The sentence with increase weightage can be considered a negative sentiment. The authors performed Naive Bayes Classifier algorithm for classification and the K-Fold Cross Validation technique for validation.

### 2.8.2 Sentiment analysis approach using POS tagging

(Bhatt, A., Patel, A., Chheda, H. and Gawande, K., 2015) The authors proposed a new rating system for amazon based on the sentiment of the product reviews. The product reviews are categorized into product and service-based reviews based on the custom rules. The authors used two different word groups that are connected to Amazon services and products. To determine if a review is a service-oriented review or a product review, each review was compared against these word lists. They looked at the breakdown of the reviews rather than just categorizing content based on keywords. They created criteria that were based on the text's structure to distinguish between reviews that were about products or services. To achieve the classification of the Amazon reviews, they combined keyword analysis with study of the review structure. Using POS tagging and ruled-based extraction (using standard expressions), they extract each feature's sentiment from the review. The sentiment phrase of the feature is provided to a polarizer method once the sentiment feature has been extracted from the text. This method evaluates the sentiment and provides a value of +1 for words that are considered to be positive and -1 for words that are considered to be negative. The method became more accurate due to the classification of reviews and sentimental analysis, which in turn provides customers accurate reviews.

(Pandey, P. and Soni, N., 2019) The researchers suggested a method called sentiment polarity categorization and POS in this study. Instead of eliminating objective content, the main goal of the study was to extract subjective content for sentiment analysis. They gathered every sentence with at least one positive or negative word and categorized it according to its parts of speech (POS). The words in each sentence were categorized using the 8 parts of speech in English, which include verbs, pronouns, nouns, adverbs, prepositions, interjections, and conjunctions. To accomplish this, they made use of a POS tagger, a program that helps categorize words according to their semantic functions. For two reasons, POS tagging is crucial to sentiment classification: It assists in two ways: 1) filtering out words like nouns and pronouns that lack sentiment, and 2) identifying words that can be employed in multiple parts of speech.

### 2.8.3 Sentiment analysis approach using Support Vector Machine Classifier (SVM)

(Haque, T.U., Saber, N.N. and Shah, F.M., 2018) This is a similar study that is performed using 48500 amazon product reviews. A supervised learning model is put forward in this study for polarizing a sizable, unlabelled product review dataset. The model makes use of two different feature extractor strategies. The Bag of Words method was used by the researchers in this study to obtain their set of features for sentiment analysis. They chose nouns and adjectives from this tagging because they are frequently the most informative for sentiment analysis. Then, a "bag" of suitable phrases was made using the selected phrases. They used the TF-IDF (Term Frequency-Inverse Document Frequency) method to find the frequently occurring words in the sentences. The word is rare and has greater significance in the document if it receives a high TF\*IDF score. A lower score, on the other hand, indicates a more common and ineffective word. The IDF measures a word's importance across all the documents in the collection, while the TF indicates how frequently a word appears in a particular document. The research employed various machine learning techniques, including Naive Bayesian, Support Vector Machine Classifier (SVC), Stochastic Gradient Descent (SGD), Linear Regression (LR), Random Forest, and Decision Tree. These methods were utilized to facilitate a comparison across varying data quantities. They conducted several simulations using cross validation, training-testing ratios, and other feature extraction processes. The findings were encouraging. In the majority of situations, 10-fold improved accuracy, while Support Vector Machine (SVM) produced the best classification outcomes.

### 2.8.4 Sentiment analysis approach using Latent Dirichlet Allocation

A unique approach for sentiment analysis utilizing Latent Dirichlet Allocation (LDA) and a Topic/Document/Sentence (TDS) model is proposed by Farkhod. To give a more accurate representation of sentiment over subjects, documents, and words, the TDS model integrates exact topic and document discovery (MDPI. (2021). The study highlights the value of sentiment analysis in comprehending consumer comments and perspectives, particularly in the context of online reviews and social media data. The suggested strategy tries to improve sentiment analysis' accuracy by including subject and document information in the analysis. The joint sentiment topic (JST) and LDA topic modelling methods that are the basis of the TDS model are intended to capture the links between topics, documents, and sentences, producing a more accurate representation of sentiment in the data. The study's experimental findings show that the TDS model performs better in terms of accuracy and precision than other approaches to sentiment analysis. The results of this study make a contribution to the field of sentiment analysis by outlining a novel method for more precisely representing sentiment across subjects, documents, and phrases 1 by combining LDA and the TDS model. The suggested method could be used in a variety of fields, such as social media monitoring and customer feedback analysis (Farkhod et al. (2021).

## 2.9 Research gaps

A research gap exists in the analysis of Amazon product reviews, particularly in focusing on specific aspects of products customer reviews. Although the researchers have researched the overall sentiments expressed by customers in reviews, they haven't given much attention to categorizing the comments. The classification and labelling of Amazon customer reviews, which include both service- and product-related reviews, aids in the development of a parallel review system. Effective sentiment extraction from reviews can be achieved with the use of the Vader tool. Topic modelling can also be used, in addition to sentiment analysis, to identify the primary themes or topics contained in a group of reviews. Even if these strategies were employed separately in several studies, they were not all used at the same time. The approaches listed above are those that are frequently used by various researchers in their studies.

## 2.10 Conclusion

In conclusion, the combination of topic modelling and VADER sentiment analysis offers a thorough and insightful method for examining Amazon product reviews. Businesses can assess the overall fulfilment or disagreement associated with their products by using VADER's ability to determine sentiment polarity, which gives a clear knowledge of customer attitudes in general. In addition, by using topic modeming approaches, we may go more deeply into the fundamental themes and topics inside the reviews, providing deep insights beyond sentiment scores. The total analysis process is improved by the combination of topic modeming with VADER sentiment analysis. While topic modelling contextualizes these sentiments by relating them to specific subjects, VADER captures the sentiment aspect of reviews. This comprehensive method enables businesses to not only understand sentiment but also to discover practical information about the causes of sentiment score.

|  |  |
| --- | --- |
| **Chapter 3** | **Research Methodology** |

# | Research Methodology

## 3.1 Introduction

Based on earlier attempts, it has been determined that there is a significant correlation between online customer buying behaviour and product reviews. My objective is to separate the reviews on Amazon into those for products and services. It is possible to develop a parallel review system using the sentiment generated by Vader algorithm for the categorised reviews. By employing the Naive Bayes technique, the accuracy of the sentiment analysis can be verified.

No previous papers have studied the possibility of combing the categorization of service and product reviews with sentiment Hybrid approach. My work provides an interesting addition to this field by combining and improving on the weak areas of earlier efforts. It emphasizes a sound methodology for data gathering and processing while additionally considering the classification of reviews. This article describes a personalized method for preprocessing and performing sentiment analysis on tweets and news articles. VADER, which is rated as the best tool for social media blogging when compared to 5 other sentiment analysis methodologies, which are used to perform sentiment analysis on text.

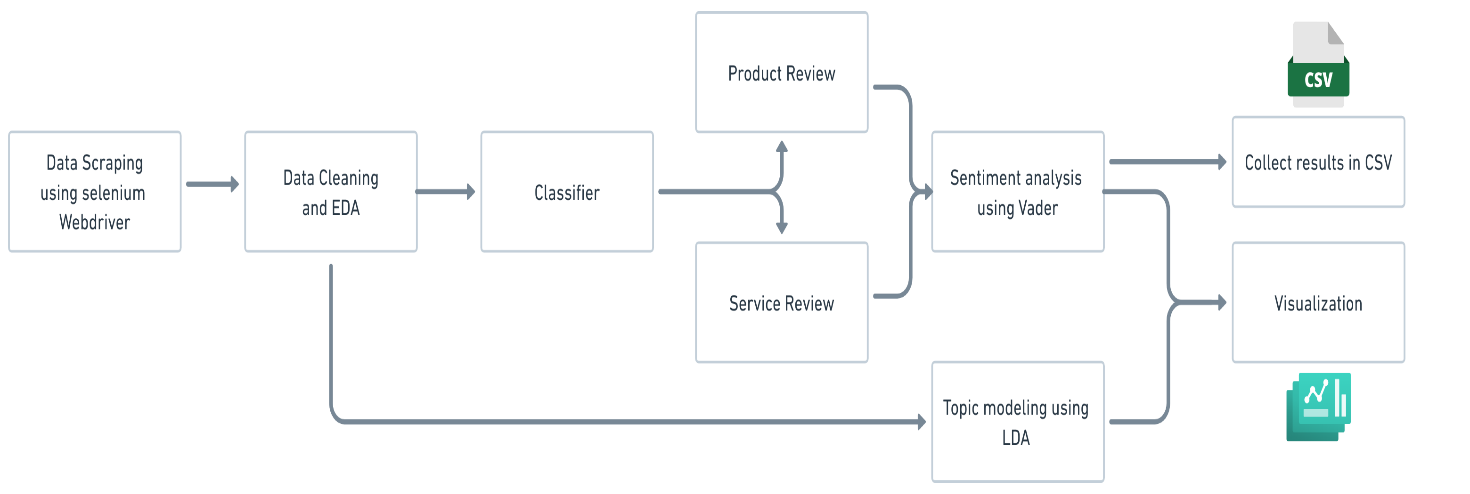
## 3.2 Business understanding

Customer feedback has become a key source of practical business insights in the modern corporate environment. The goal of this section is to unearth strategic insights that facilitate informed decision-making by exploring into the crucial function of business analysis in the context of Amazon review data. Businesses can improve their services, enhance customer experiences, and gain a competitive edge in the market by looking at the customer attitudes and opinions detailed in reviews.

The conclusions drawn from sentiment analysis and topic modelling set the groundwork for practical strategies rather than just offering a theoretical understanding. Businesses can use the sentiment ratings to determine how customers feel generally about their goods and services. While negative attitudes might help to enhance services, positive sentiments can be used for marketing activities. Businesses can use the highlighted subjects as a lens through which they can align their strategies with those of their customers. Businesses might give priority to improving their customer service channels if, for instance, frequent trends revolve around "customer support responsiveness." These opinions are crucial for developing customer-focused, data-driven company strategies that increase profitability and customer loyalty. Analysis of feelings provide insights about particular items as well as general brand impressions. Positive feelings are associated with consumer loyalty and satisfaction, which ultimately helps build a solid brand reputation. Additionally, keeping track of how sentiment shifts over time enables organizations to evaluate the success of their strategy and make necessary adjustments. A complete perspective is provided by the special combination of sentiment analysis and subject modelling, boosting the competitive edge. In the below section 3.3 we a detailed and plan is formulated to develop methods to uncover the sentiments and insights behind the review text.

## 3.3 Proposed methodology

Based on the research gaps found in the current studies, it is evident that a new methodology is required to extract the insights behind Amazon product reviews. The proposed methodology module offers a step-by-step guide for undertaking a thorough examination of sentiment and subject trends in service and product review data. This systematic method lays out the sequential procedures and methods that will be used to gather useful insights from the dataset, highlighting customer opinions, preferences, and concerns. This methodology incorporates topic modelling and sentiment analysis in order to reveal hidden patterns and provide a deeper knowledge of the sentiments expressed and the topics covered in service and product reviews.

Figure **2**, Proposed methodology flow chart

Based on the Figure 2 The first stage is to get the necessary Amazon product review data for Amazon.com using the Selenium Web driver, and store it in a CSV file. Later, the data is cleaned and transformed using pre-processing methods. Exploratory data analysis (EDA) is performed to get an insight on the extracted data. The cleaned data is used to perform feature selection and to identify target attribute for modelling. The cleansed data is then fed into a classifier module to categorise the reviews into product and service reviews. The categorised data is given as an input to the lexicon-based model VADER for sentiment analysis. As a result, the input data are processed by VADER's sentiment analyser, which assigns sentiment polarities to each of the reviews. The cleaned data is then given into LDA to do topic modelling and find the most popular product-related phrases within each topic.

The methodology is composed of the following modules:

1. Data Collection
2. Data Pre-processing
3. Exploratory data analysis (EDA)
4. Classification of reviews into product and service reviews
5. Sentiment analysis using Vader
6. Topic modelling with LDA

## 3.4 Data collection

It is clear from the literature review's conclusion how product reviews might affect consumers' purchasing decisions. We need a dataset in order to conduct further analysis and develop a new review system based on review sentiments. There are several ways to extract data from amazon.com such as using third party web scrapping tools, Amazon webservice API, python libraries and Web scrapping using selenium. Each of the methods have its own benefits and downfalls.

By initiating HTTP queries, processing the HTML text, and extracting the needed data. Web scraping uses Python libraries to gather data from websites. The navigation, search, and modification of the parse tree are performed using Beautiful Soup, an effective tool for parsing HTML and XML documents. Amazon blocked access to the Amazon website's data extraction when utilizing this method. The request was responded to with a 403-access-denied response code from Amazon.com. To prevent getting too many queries, Amazon limits the usage of automated services to submit API calls. Data collection was aided by the usage of the third-party online scraping application known as web scrapper 2020. The tool could switch IP addresses after each API request; however, AWS (Amazon webservice) recognized it as a bot. This function of the Amazon.com website blocks sending additional API requests to the page. This approach was likewise unsuccessful in obtaining the required information from the website because we are unable to send any requests. An official service provided by Amazon for sellers to track sales and product reviews is the Amazon webservice advertising API. This API offers the ability to make requests to a website in order to collect data in JSON format. The seller can use the data that was extracted to enhance customer service or sales. An authorization from Amazon was required to use this API service. According on the analysis and intended use of the data, access is given.

It is clear from the previously mentioned options; selenium web crawling is the most effective method for obtaining Amazon reviews of products. Selenium is a graphical web automation tool that enables testers to swiftly finish partial case tests using a mouse and conveniently record browser actions. Additionally, Beautiful Soup's function is used to parse data from XML and HTML files. Selenium uses simulations of human behaviour to navigate websites and gather data. This makes it easier to go around Amazon's bot detection technology (Huang and Li, 2019). Our own data requirements were used to design the

framework for data extraction. As a result, it is simple to extract the data in the precise format that we needed.

We made the decision to pull data from smartphones made by brands like Apple, Samsung, Xiaomi, Realme, and One Plus for this investigation. Since there are more than a dozen phones sold by these brands on Amazon, a selection criterion was necessary. A product selection criterion is established for data extraction following thorough consideration of the study objective.

### 3.4.1 Data selection criteria

1. Products need to be categorized under names like Apple, Samsung, Xiaomi, Realme, and One Plus.

2. The pricing of the product should range from $429 to $2000.

3. Product release date should be in between 2021 and 2023.

4. Products with comparable features should be taken into account. It may be utilized subsequently for brand competitor analysis.

5. There should be more than 100 Amazon product reviews for the chosen item.

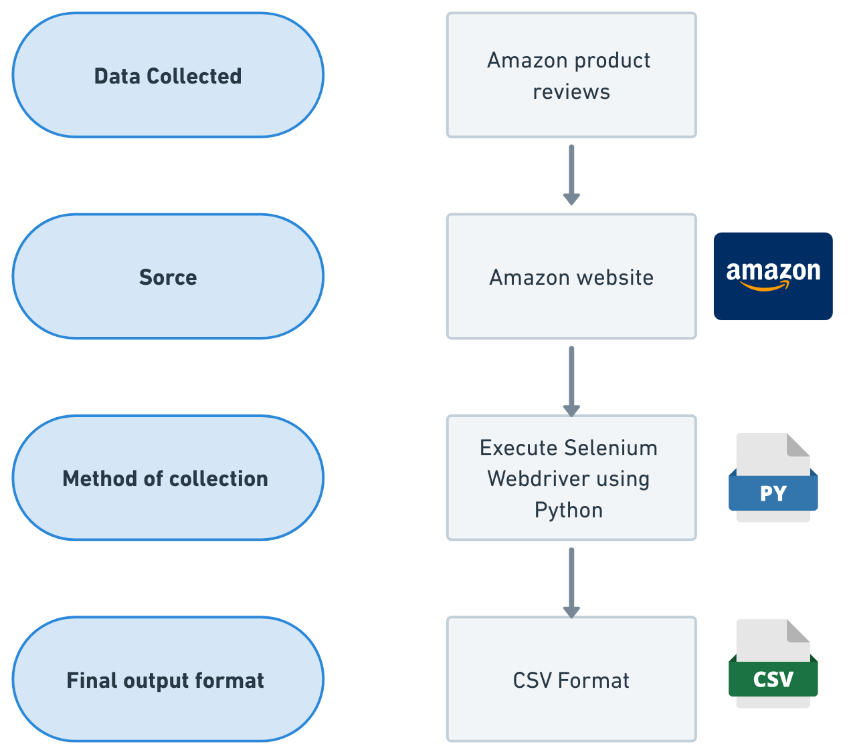


Figure 3,Data collection procedure

**Figure 3 shows the data collection procedure in high level. Based on the above data selection criteria amazon product reviews are extracted from amazon websites using Selenium web driver. The extracted data are then saved as a CSV format.**

**In Figure 4 the detailed flow chart of the data extraction procedure is given. Based on the selection criteria, 20 phones from different brands are selected. Amazon website URL of the respective phones are saved in a configuration CSV file. Each URL is passed to the website to access its product reviews page. From the HTML tags the data is extracted using python code with the help of XPATH. Amazon holds only 10 product reviews per page. To access all the reviews each review page should be visited and data should be extracted. The next page navigation will be continued till the last page. Once the last page is reached, the python code will search for next URL from the config file. Like this the product reviews of all the products are extracted. Once all the product reviews are extracted the data is saved as a CSV file format.**

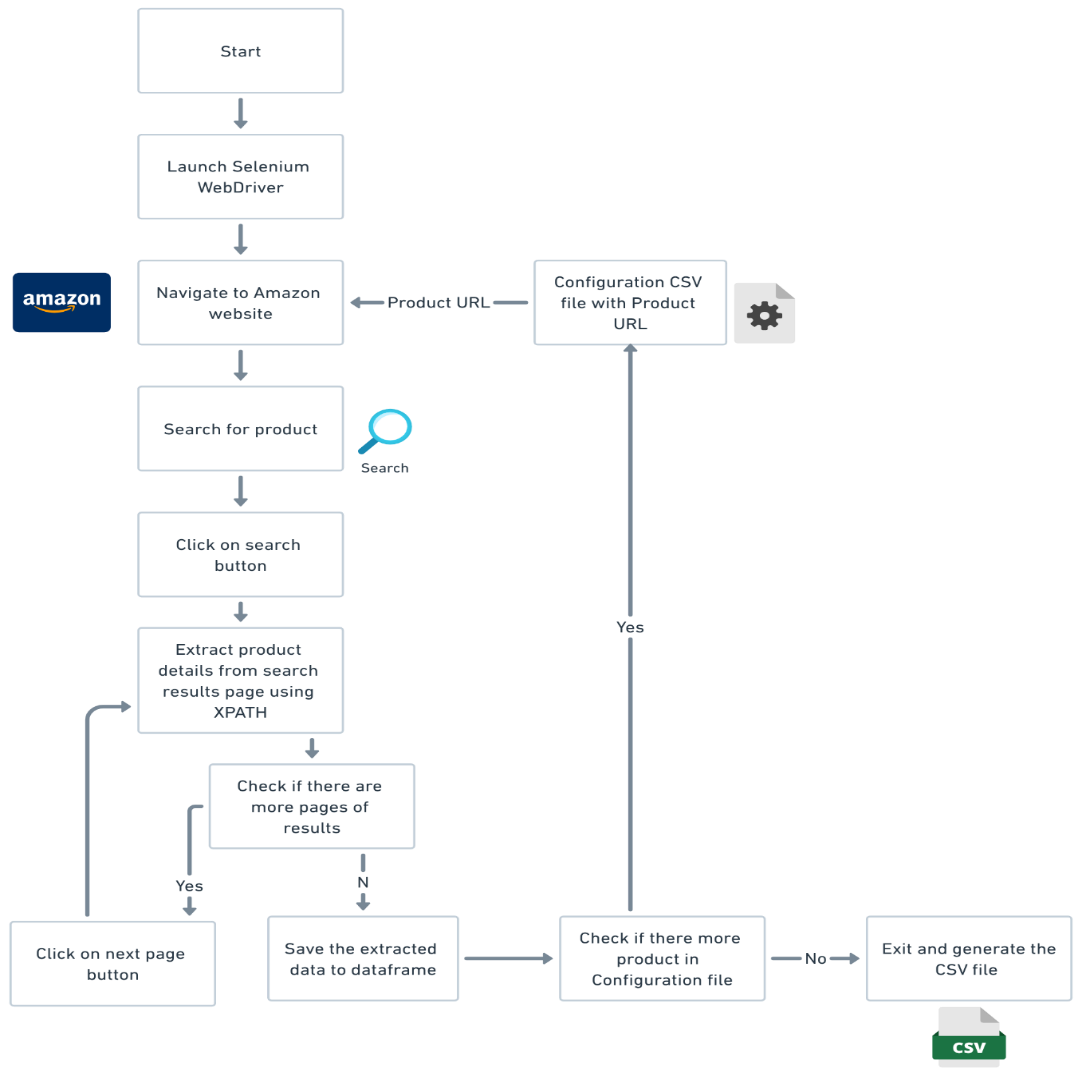


Figure 4,Data extraction procedure, detailed flowchart

### 3.4.2 Dataset description

**This dataset provides comprehensive information about product reviews from various companies, including the overall product ratings, review counts, review titles, reviewer names, review ratings, review text, and helpfulness ratings. The dataset aims to facilitate sentiment analysis, customer feedback analysis, and other related research in the field of product reviews and customer satisfaction.**

df = pd.read\_csv('D:/amazon\_review\_data/Amazon\_reviews.csv', encoding=encoding)

print(" \nBefore removing anomalies from DataFrame : \n\n",

df.isnull().sum())

print("Total number of rows = ",len(df))

df.sample(n = 50)

print(df.describe())

**Python code for reading the csv file and to check for anomalies**

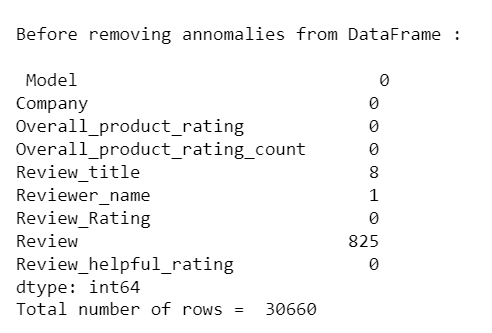
****

Figure 5,Dataset description

**The python code shown above is used to read the csv file and print the description of the dataset. By analysing the output shown in the figure 5 (section 3.3.2), it is clear that null values are present in the dataset. In the figure 6 a snapshot of the dataset is provided with nine attributes.**

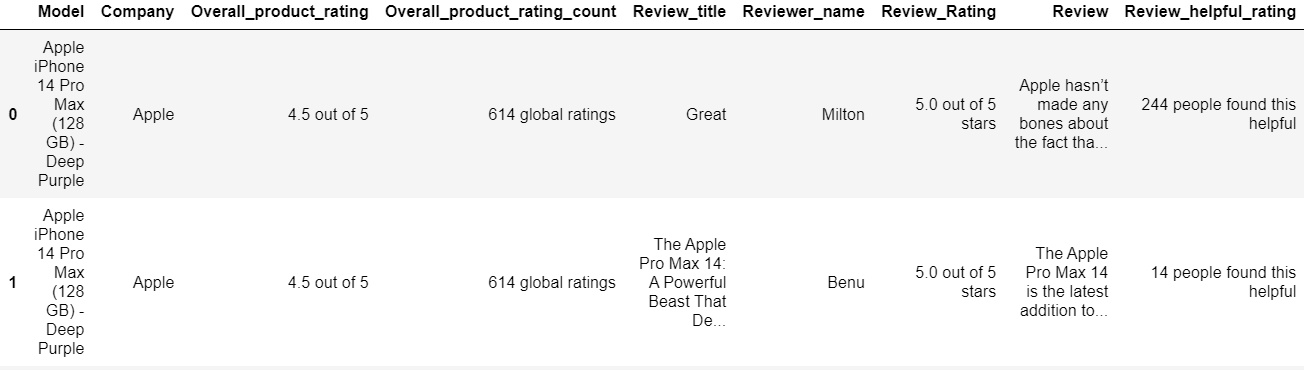
****

Figure 6,Dataset sample

**The dataset consists of 9 attributes with 30661 rows.**

**Below are the attributes available in the dataset:**

* **Model: The unique identifier or name of the product model being reviewed.**
* **Company: The name of the company producing the product being reviewed.**
* **Overall\_product\_rating: The overall rating of the product, typically on a scale of 1 to 5, representing the average rating given by all reviewers.**
* **Overall\_product\_rating\_count: The total count of ratings submitted for the product.**
* **Review\_title: A concise title or headline for the review provided by the reviewer**
* **Reviewer\_name: The name or username of the individual who submitted the review.**
* **Review\_Rating: The specific rating provided by the reviewer for the product, usually on a scale of 1 to 5.**
* **Review: The detailed text content of the review where the reviewer shares their opinions, experiences, and feedback about the product.**
* **Review\_helpful\_rating: The number of times other users found the review helpful, indicating its usefulness within the community.**

**This dataset aims to be a valuable resource for businesses, researchers, and analysts interested in understanding customer perceptions, identifying product strengths and weaknesses, and studying the impact of reviews on product sales and brand reputation. Researchers can leverage this dataset to develop sentiment analysis models, customer behaviour studies, and recommendation systems to improve product quality and enhance user experiences.**

## 3.5 Data Preprocessing

Data preparation is essential for turning unorganized, unclean data into a clean, logical dataset that can be used for data analysis and machine learning. Raw data frequently contains flaws, contradictions, and noise that can reduce the precision and efficiency of future analysis or machine learning models. In order to ensure that the data is usable for additional analysis or modelling, data preparation entails a number of stages and strategies aimed at addressing these problems.

In the literature review we discussed several methods to clean the data to achieve accurate sentiment analysis. The VADER (Valence Aware Dictionary and Sentiment Reasoner) sentiment analysis tool does not particularly require a deep data pre-processing. However, it can still be beneficial to do so before doing sentiment analysis. On the other hand, Topic modelling using Latent Dirichlet Allocation (LDA) requires preprocessing before execution. Preprocessing ensures that the text is in a suitable format for analysis, reduces noise, and enhances the effectiveness of the LDA algorithm. After carefully examining the raw data, an efficient preprocessing method is chosen based on the input requirements of Vader and LDA.

### 3.5.1 Basic data preprocessing steps.

Based on the initial analysis of the raw data few anomalies were found. These irregularities should be rectified before implementing a deeper cleaning procedure.

1. **Remove duplicate data:**

The analysis found that out of 30660 row, 21 rows are duplicates. To remove duplicate rows from the Data Frame, we can use the ‘*drop\_duplicates ()*’ function. This function removes duplicates based on the values in all columns by default.

cleaned\_df = cleaned\_df.drop\_duplicates()

Python code to remove duplicates

1. **Remove Null values from Dataset:**

There are 823 null values in the Review column, 1 null value in the Reviewer\_name column, and 6 null values in the Review\_title field. To remove rows with null values from your Data Frame, you can use the ‘*dropna ()*’ function. By default, this function removes rows that have at least one null value in any column.

cleaned\_df = cleaned\_df.dropna()

cleaned\_df = cleaned\_df.reset\_index(drop=True)

Python code to remove rows with null value and reset index

1. **Fill empty fields with zeros:**

Review\_helpful\_rating has 19151 empty fields that can be filled with zeros. No one found the review useful is how zero values are expressed**.**

1. **Trim the rating values to numeric only values:**

Review rating columns with alphabetic letters include Overall\_product\_rating, overall\_product\_rating\_count, and Review\_Rating. To effectively do an exploratory analysis, this should be taken out.

for a in range (0,len(cleaned\_df['Overall\_product\_rating'])):

cleaned\_df['Overall\_product\_rating'][a] = cleaned\_df['Overall\_product\_rating'][a].replace(' out of 5', '').strip(

for b in range (0,len(cleaned\_df['Overall\_product\_rating\_count cleaned\_df['Overall\_product\_rating\_count'][b] = cleaned\_df['Overall\_product\_rating\_count'][b].replace(' global ratings','').strip()

if ',' in cleaned\_df['Overall\_product\_rating\_count'][b]:

cleaned\_df['Overall\_product\_rating\_count'][b] = cleaned\_df['Overall\_product\_rating\_count'][b].replace(',', '').strip()

for c in range (0,len(cleaned\_df['Review\_Rating'])):

ceaned\_df['Review\_Rating'][c] = cleaned\_df['Review\_Rating'][c].replace(' out of 5 stars', '').strip()

for d in range (0,len(cleaned\_df['Review\_helpful\_rating'])):

cleaned\_df['Review\_helpful\_rating'][d] = cleaned\_df['Review\_helpful\_rating'][d].replace(' people found this helpful', '').strip()

if cleaned\_df['Review\_helpful\_rating'][d] == "One person found this helpful":

cleaned\_df['Review\_helpful\_rating'][d] = cleaned\_df['Review\_helpful\_rating'][d].replace('One person found this helpful', '1').strip()

Python code to trim the values in below column

* Example from “*Overall\_product\_rating*” attribute: One of the values in Overall\_product\_rating attribute value in raw data is, “4.5 out of 5”. We only need the original rating “4.5” for the analysis, rest of the text can be removed.
* Example from “*Overall\_product\_rating\_count*” attribute: One of the values in Overall\_product\_rating\_count attribute value in raw data is, “614 global ratings”. We only need the original rating “614” for the analysis, rest of the text can be removed.
* Example from “*Review\_Rating*” attribute: One of the values in Review\_Ratingattribute value in raw data is, “5.0 out of 5 stars”. We only need the original rating “5.0” for the analysis, rest of the text can be removed.
* Example from “Review\_helpful\_rating” attribute: One of the values in Review\_helpful\_rating attribute value in raw data is, “2 people found this helpful”. We only need the original rating “2” for the analysis, rest of the text can be removed.

1. **Combine the data in “Review\_title” attribute and “Review” attribute:**

According to the study, the review's title and content are more comparable than previously thought. Additionally, the review's title offers useful details that may affect the tone of the text. Utilizing the Vader tool, data from both attributes are blended to provide the sentiment analysis with maximum coverage.

1. **Translate non-English reviews to English using Google API translator python library:**

This library allows you to easily integrate translation functionality into your Python applications. We can now use the ‘*translate ()’* method of the translator object to perform text translation.

### 3.5.2 Deeper data preprocessing steps.

When using the data for topic modelling using LDA, additional stages of data cleaning were taken in addition to the baseline cleaning. Changing the product review to lowercase will prevent the algorithm from misinterpreting the words "Mobile" and "mobile" as the same. Following that, punctuation and numbers should be removed to further clean up the text. Punctuations or numbers don’t hold any sentiment information. The following step is to eliminate any double spacings that may have resulted from deleting the punctuation and the numbers. Stop words are eliminated by importing them from the nltk package, which transforms the product review text form from a string to a list of words. Stop words provide little information and may make sentiment difficult to interpret. The words in the list were stemmed to remove the end words so that the algorithm would detect related words as the same words. In addition to that URL, ’Mentions’ and hashtags are removed using regular expression support module available in python library.

### 3.6 Exploratory Data analysis (EDA)

Exploratory Data Analysis (EDA) helps the finding of hidden patterns, correlations, and features within data and acts as a fundamental step in understanding and obtaining insights from datasets (Cohen 2013). EDA involves finding repetitive patterns and connections in the data. Seasonal variation in time series data, groups of related observations, or correlations between variables are a few examples of patterns (Chambers et al., 1983).

**N-grams: Unveiling Linguistic Patterns in Textual Data**

N-grams are a key concept in Natural Language Processing (NLP) and are crucial for understanding and gathering contextual data from text input. Nearby sequences of N items, each of which can be a word, character, or symbol, are known as N-grams (Jurafsky & Martin, 2019). They are useful tools for identifying linguistic patterns and associations between recurrence in a text.

N-grams are divided into groups according on how many components they include. Individual words or characters are represented by unigrams, or 1-grams, while pairs of words or characters are represented by bigrams, or 2-grams, and sequences of three consecutive words or characters are described by trigrams, or 3-grams. This idea is expanded to longer sequences by higher-order N-grams (Jurafsky & Martin, 2019). These N-grams contribute to an enhanced understanding of linguistic context by revealing information about word or character frequency as well as language structure.

In Figure 6 (section 3.5), shows that top 10 unigrams present in the dataset. From the figure it’s evident that ‘phone’, ’good’, ’camera’ are some of the frequently used words. To get a better context of the text used, bigrams and trigram graph need to be analysed

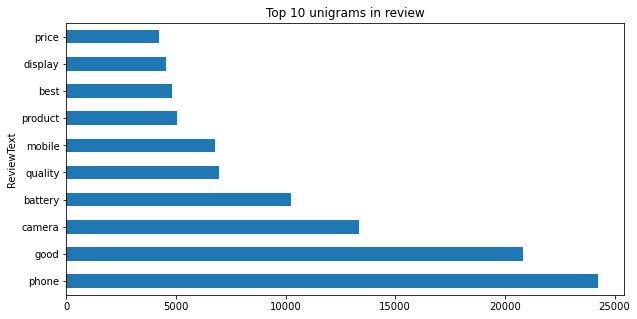


Figure 7,Unigram graph based on amazon reviews

In figure 7 (Section 3.5), top 20 bigrams are given. This gives a better context to the wordform this graph it is visible that most of the people from the overall reviews are giving positive reviews. At the same time issues such as ‘heating’, ‘price range’ of the phones are also discussed in the review text. The bigram provides a meaningful insight when compared to the unigram graphs

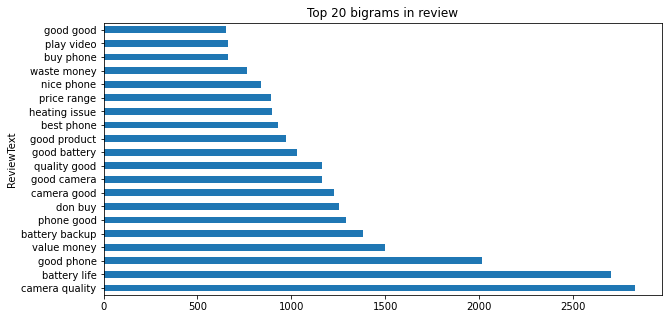


Figure 8,bigram graph based on amazon reviews

In figure 8 (section 3.5), top 20 trigrams are given. This gives a better context to the wordform this graph it is visible that most of the people from the overall reviews are giving positive reviews. Most of the reviews are discussing about the features such as battery, camera finger print sensors and so on. The Unigram and bigram graph will not provide a clear picture of the most discussed words. Thats why we are using Trigram to visualize the most discussed 3 words in better context.

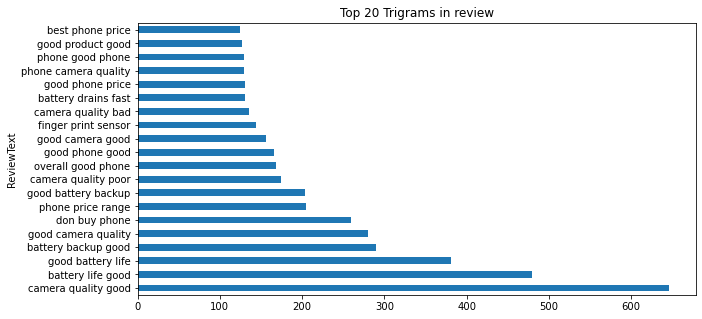


Figure 9,Trigram graph based on amazon reviews

**Review linguistic features**

In Figure 9 (section 3.5) shows the complexity of the reviews sentences for each of the companies in the review text. he total number of characters gives you an idea of how long the text is. Longer texts might be more detailed or in-depth, while shorter texts could be concise or to the point. Very long texts may indicate complex subjects or detailed explanations, while shorter texts might be summaries or overviews. From the graph it is clear that Apple and Xiaomi customers tend to write in depth reviews. At the same time OnePlus and Relme customers tried to present their reviews in a surmised version.

A higher word count often suggests more detailed content. Longer texts can explore topics more thoroughly and provide comprehensive information. Shorter texts may be more easily digestible and suitable for quick consumption, while longer texts may require more time and attention. From the figure 9 we can see that Apple and Xiaomi users writes their reviews in detail. On the other hand, Samsung, OnePlus and Realme users keeps the review short and crisp.

High word density (more words in a given space) can indicate that the text is focused and contains relevant information. However, excessively high word density might make the text dense and harder to read. Low word density might suggest that the text is more verbose and less concise. It could also mean that the text is repetitive or lacks detail. From the figure it’s evident that all the brand users write their reviews in detail and without any repetitions of content.

The number and types of punctuations (commas, periods, exclamation marks, etc.) can reveal the complexity and variety of sentence structures in the text. punctuation usage can convey the tone of the text (formal, casual, enthusiastic) and emphasize certain points or ideas. Among the 5 brand users, Xiaomi, Apple and OnePlus users tend to use the punctuations proper and keep their reviews in proper sentence structures. Samsung and Realme users are poor in maintaining a good sentence syntax.

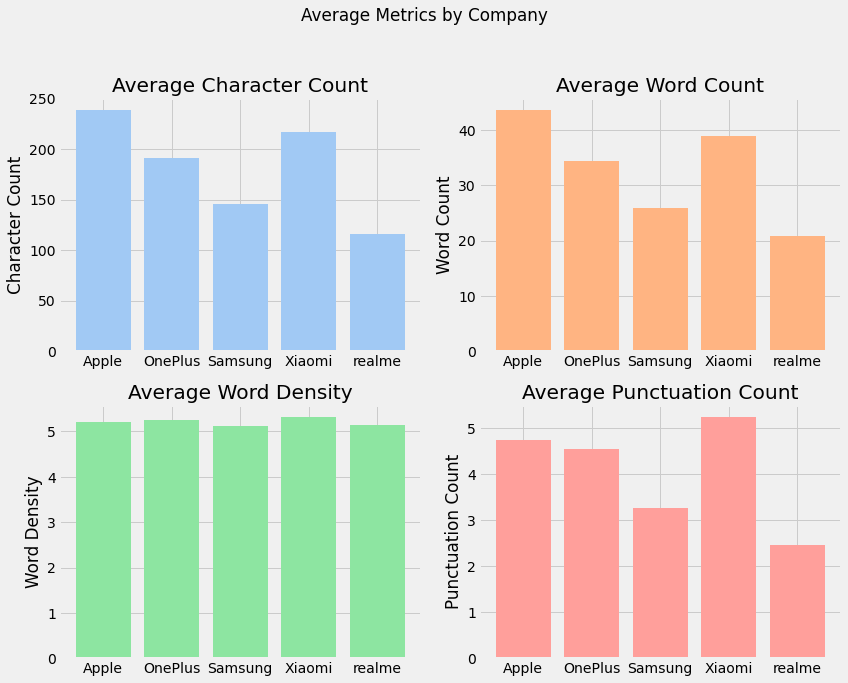


Figure 10, Character count, word count, word density, punctuation count of amazon reviews

In figure 10 (section 3.5) most of the customers have given the rating 5.0,4.0 and 1.0. The desire for impact is a significant factor that can influence consumers to choose extreme ratings (5 or 1) in product reviews. This behaviour is driven by the understanding that extreme ratings have a more pronounced effect on the overall average. Consumers who feel strongly about a product's quality or their experience may choose an extreme rating in an effort to amplify their impact. A single 1-star or 5-star rating can significantly shift the overall average, attracting attention to their review. They might want to warn others about a poor experience (1-star) or enthusiastically recommend a product they found excellent (5-star). Similar to the factors that influence people's decisions to post reviews for various goods and services, most customers only want to leave a review when they have strong positive or negative feelings about its Nakayama, M. and Wan, Y., 2019.



Figure 11, Overall review rating for each company

Figure 11 (section 3.5) show the total review given for each of the products in amazon. In other words, we can say that the models with the greatest number of ratings are the ones which are popular in amazon sales. There are several factors that affects the sale in amazon for a particular bard or model. Advertisements are one the major factor among them. Other factors such as word of mouth, social media influence, and brand perception can affect the customer purchasing behaviour. From the graph it is clear that the most popular brand in amazon is OnePlus, Samsung and Xiaomi. Even though Apple is a leading brand customer are not incline to buy any of the Apple phones from Amazon website. This Maily happens because Apple company tend to sell their product through Apple online store or physical stores.

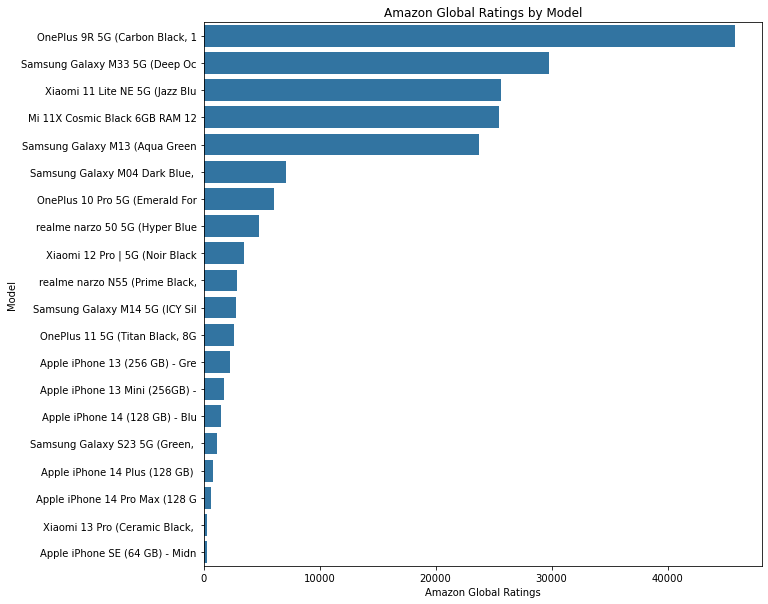


Figure 12, Total ratings given for each model

In figure 12 (section 3.5) average helpful review is given, among the models the reviews written by Apple SE users were found more use full by peer customer. The number of average helpful rating is 60, which is very high when compared to other models available in the market. This means that Apple SE reviews were red by other customers more when compared to revies of the different models available in the market.

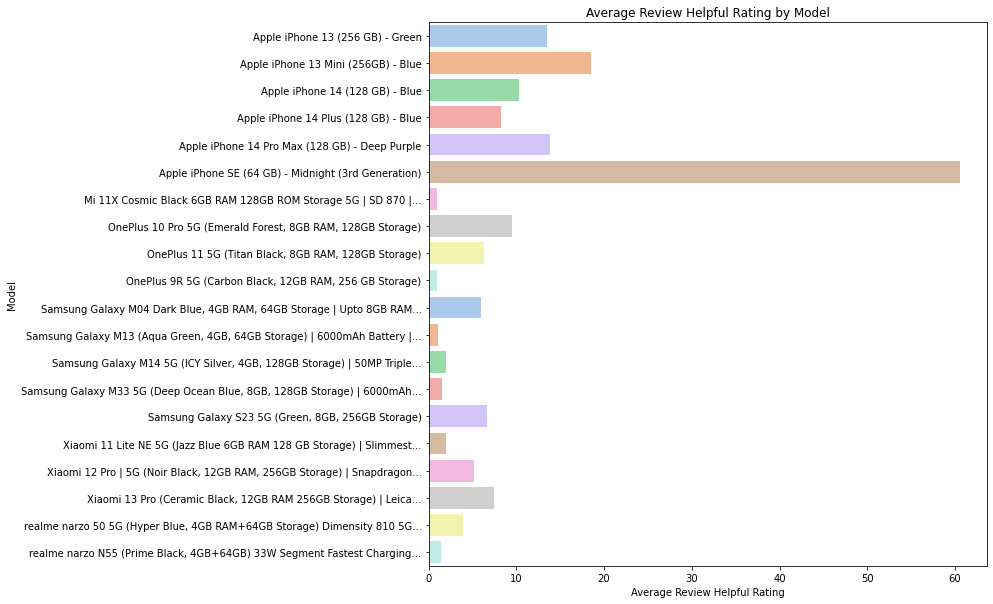


Figure 13, Average number of helpful ratings for each model

In figure 13 (section 3.5) shows the total count of emoticons in each smartphone models in the dataset. The use of emoticons in text, including product reviews written by customers, can convey various emotions, attitudes, and sentiments in a concise and visually appealing manner. Emoticons are small graphical symbols or icons that represent feelings, expressions, or concepts. From the figure 13 it’s clear that Xiaomi, Samsung and one plus customers tend to use more emoticon in their review text.

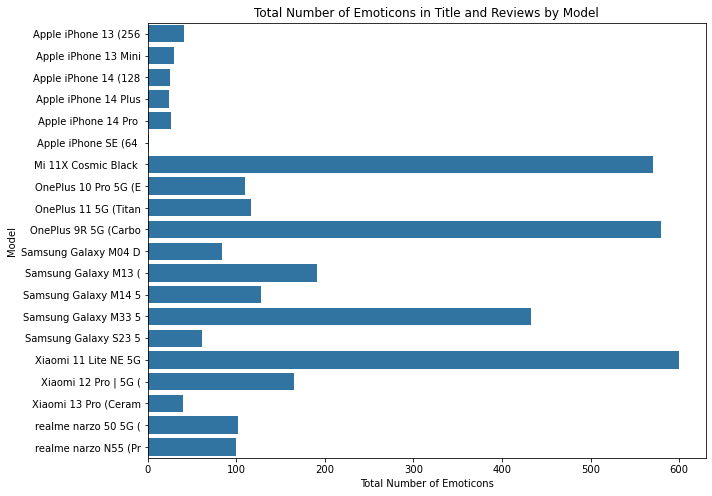


Figure 14, Total number of emoticons in reviews for each model

Word clouds can be particularly useful for visualizing the most frequently mentioned words in product reviews, customer feedback, social media comments, or any other text-based data. They provide a quick and intuitive way to identify prominent themes and sentiments within the text. From the figure 15 (section 3.5) it is clear that the most discussed product features in reviews are related to battery, camera and display of the phone, Mobile, Good, Samsung. These words are the most frequently used phrases in the reviews. Since these words are high in number, they can have a great impact on overall sentiment of the reviews.

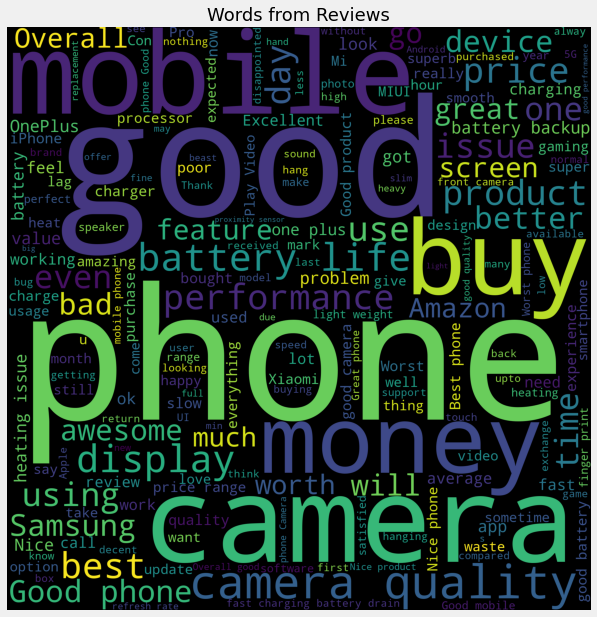


Figure 15, Word cloud for Overall reviews

## 3.7 Classification of reviews into product and service reviews

Reviews on Amazon can be broadly categorized into two different types: service reviews and product reviews. Reviews of services typically focus on how customers felt during the ordering process, order fulfilment, delivery, and customer service interactions. These reviews frequently concentrate on elements like delivery time, packing quality, and the overall shopping experience. The actual features, functionality, and quality of the purchased item are the focus of product reviews, in contrast. The functionality, durability, and any problems customers have with the product are discussed by customers. Differentiating between service and product reviews allows businesses and consumers to learn insightful information about the actual purchasing process as well as the actual features of the products, allowing informed decision-making and improving general satisfaction among consumers.

(Bhatt, A., Patel, A., Chheda, H. and Gawande, K., 2015) Based on the studies done by Bhatt and his team, they proposed an algorithm to classify the amazon reviews into product and service reviews. In light of their work, we can classify things with a few modest adjustments and considerably more effectively. Two sets of words related to Amazon services and products were used by the authors. Each review is compared to this list of words to determine whether it is a service-oriented review or a product review. In addition to using keywords to identify reviews, they also looked at the reviews' structures. To categorize reviews as either service-based or product-based, they developed criteria based on the text's structure. In their strategy, they used keyword analysis with review structure analysis to achieve the classification of the Amazon reviews.

The text structuring criteria is as follows:

*“a. Perform a service review test where the review is tested for occurrence of service words, i.e., if the review length is shorter than 15 words and service words are found in the review. The review is classified as service review else if the length of review is greater than 15 then more than 2 service must occur in the review for it to be a service review.*

*b. If the review fails for the service test, then it is tested for features of a product (such as camera, microphone and battery) if these exist then we classify the review as a feature review.”* (Bhatt, A., Patel, A., Chheda, H. and Gawande, K., 2015).

# Lists of service-related words and feature words

service\_keywords = ["service", "staff", "efficient", "delivery","delivered","amazon",

"customer service","replacement","service center","packing","packaging","arrived",

"refund","Shipping","return process","warranty","assistance","complaint",

"live chat","chat support","helpdesk","support team","packed","delivery boy","good condition","appario","secure package","package","customer care","cs team","delivery experience","delivery agent","exchange","damaged","shipment","box"]

product\_keywords = ["samsung","oneplus","apple","xiaomi","camera", "microphone", "battery","display","fingerprint","sensor","charging","charge","performance","processor", "ram","network","bluetooth","gps","upgrade","cam","screen","photography","speaker","android","selfie","heat","hang","internet","game","gaming","brand","finger print",

"Face unlock","5g","4g","video"]

Python code, Extra added keywords

In this section 3.6 the aim is to explain how we are labelling the amazon reviews as product and service reviews. In the previous studies conducted by Bhatt, we have seen several methods to categorize the review as product review and service review. In our method we are implementing additional rules based on the amazon review in our dataset.

Algorithm rules implemented:

1. If the number of product related words available in the sentence is greater than the number of service-related words, the review can be categorised into product review
2. If the number of service-related words available in the sentence is greater than the number of product-related words, the review can be categorised into service review
3. If both the conditions are failed the algorithm will check, if the sentence length is less than 15 words and contains any service words. Then the review is labelled as service review.
4. If the above condition is also failed, the algorithm will check if the sentence length is greater than 15 words and contains any service words. Then the review is labelled as service review.
5. If all these checks are failed, the review is labelled as product review.

## 3.8 Sentiment analysis modelling using Vader

Sentiment analysis is a method that involves assessing the subjectivity and polarity of text by taking into account the impact of words and the emotional meaning of words. Different techniques to evaluate sentiments are included in this procedure, often known as sentiment analysis. Lexicon-based methods and machine-learning-based methods are the two main techniques for automatic sentiment extraction. Our review of the literature reveals a general preference for the VADER lexicon-based approach, which is especially appropriate for social media blogging. This discovery coincides with the study of C.J. Hutto and Eric Gilbert done in 2014.

In sentiment analysis, the polarity (positive, negative, or neutral) and intensity of the emotional tone expressed in a text are frequently evaluated using sentiment scores, which are numerical representations of the emotional tone. Sentiment scores, which fall within a specific range, express the strength of the sentiment, with higher values signifying stronger feelings. While context and the words around the score play a key role in appropriate interpretation, the sign of the score shows whether the feeling is positive or negative. Tools like VADER manage complexities like negation and context when calculating sentiment scores from the text's sentiment-heavy words and phrases. Setting threshold values establishes the boundaries between these groups when categorizing VADER sentiment ratings into positive, neutral, and negative labels. With a scale of -1 (most negative) to +1 (most positive), neutrality is represented by a score of 0. Although this method makes sentiment categorization simpler, it might miss complexities. You can think about employing a finer-grained strategy that incorporates additional sentiment categories for deeper categorization.

The threshold of deeper categorization can be done based on the Vader sore as below. The threshold values for each sentiment category based on the sentiment score range.

• Positive: Score > = 0.5

• Negative Score <= -0.5

• else neutral: Score

# init the sentiment analyzer

sia = SentimentIntensityAnalyzer()

Python code to initialize the Vader sentiment analysis package

Dataset can be categorised based on the Vader sentiment score; this helps us to identify the sentiment of a review in a deeper sense. To perform sentiment analysis using VADER (Valence Aware Dictionary and sentiment Reasoner), we can use the Natural Language Toolkit (NLTK) library in Python. NLTK provides a pre-trained VADER sentiment analysis module that you can readily use for analysing text sentiment. The combined fields of review title attribute and review text is given to the Vader tool to analyse the sentiment. Once the sentiment score is found the score is categorised based on the above scores. Later a new attribute named “Sentiment” is created and labelled against the respective review text fields in dataset.

# Define a function to get sentiment score

def get\_sentiment\_score(text):

return sia.polarity\_scores(text)["compound"]

# Define a function to get polarity label

def get\_polarity\_label(score):

if score >= 0.05:

return "Positive"

elif score <= -0.05:

return "Negative"

else:

return "Neutral"

# Apply the polarity label function to the sentiment scores and store labels in another new column

cleaned\_df['Polarity'] = cleaned\_df['Sentiment\_Score'].apply(get\_polarity\_label)

Using the above python code, the sentiment label is stored into a newly generated attribute named ‘'Polarity ‘’. The Sentiment score is stored inside the attribute named ‘Sentiment score’.

## 3.9 Topic modelling

Two distinct methods for analysing text data are sentiment analysis and topic modelling, each with a particular goal and strategy. Sentiment analysis attempts to ascertain the emotional sentiment, whether good, negative, or neutral, communicated in text. Since its main objective is to evaluate sentiment polarity and intensity, it is useful for comprehending viewpoints, reviews, and emotional trends. This entails classifying text into sentiment categories using lexicon-based techniques or machine learning models, producing sentiment scores or labels as outputs. Topic modelling, on the other hand, focuses on locating hidden themes within a group of texts. Its main goal is to locate and classify related documents according to underlying content trends. In order to extract subjects, methods like Latent Dirichlet Allocation (LDA) are used. This method reveals word clusters and their distribution throughout the text. Along with the sentiment analysis, topic modelling can offer contextual information to help interpret sentiment within different thematic contexts.

To perform the topic modelling using LDA we need to extract the corpus and dictionary phrase from the reviews. The below python code is used to perform that method.

# Create a dictionary and corpus

dictionary = corpora.Dictionary(cleaned\_df['processed\_text'])

corpus = [dictionary.doc2bow(text) for text in cleaned\_df['processed\_text']]

Python code to extract corpus and dictionary from the review text

Topic models are assessed using two metrics: coherence and perplexity scores. Lower numbers indicate greater performance when measuring a model's perplexity, which is a measure of how well it predicts a held-out test set. Coherence, which assesses the interpretability of the themes, quantifies the distance between words within a topic. Better topic interpretability is indicated by a higher coherence score (Zvornicanin,2022).

### 3.9.1 Identify the optimal number of topics

# List of different numbers of topics to try

num\_topics\_list = [1, 2, 3, 4, 5]

coherence\_scores = []

# Loop through different numbers of topics

for num\_topics in num\_topics\_list:

# Train the LDA model

lda\_model\_check = LdaModel(corpus=corpus, id2word=dictionary, num\_topics=num\_topics, random\_state=42)

# Calculate coherence score

coherence\_model\_lda = CoherenceModel(model=lda\_model\_check, texts=cleaned\_df['processed\_text'], dictionary=dictionary, coherence='c\_v')

coherence\_score = coherence\_model\_lda.get\_coherence()

coherence\_scores.append(coherence\_score)

Python code to identify the optimal number of topics

To identify the optimal number of topics the coherence score is analysed for N number of topics. The above-mentioned python code is used to find coherence score of 5 topics. A line graph is plotted based on the score values.

### 3.9.2 visualise the LDA model using pyLDAvis visualisation library

The open-source Python module pyLDAvis aids in the analysis and creation of extremely interactive visualizations of the clusters produced by LDA. Its purpose is to assist users in understanding the subjects in a topic model that has been adjusted to a corpus of text data. In order to provide information for an interactive web-based visualization, the software extracts data from a fitted LDA topic model.

# Visualization using pyLDAvis

vis\_data = gensimvis.prepare(lda\_model\_build, corpus, dictionary)

pyLDAvis.display(vis\_data)

Python code to visualise the insights found using LDA model

The above code is used to create an interactive visualisation chart. This chart can be used to analyse the insight generated using the LDA model efficiently.

## 3.10 Summary

This thesis' third chapter focuses on the challenging process of managing and analysing data, covering data gathering techniques, data preparation, data visualization, and data modelling. An exhaustive data collection strategy is described, guaranteeing the collection of a diverse and wide range of dataset, in the attempt to decipher the sentiments and insights buried within Amazon reviews. The chapter described the crucial stage of data preprocessing like, like tokenization, stop-word removal, and stemming to make it more suitable for analysis. Data visualization techniques enable a detailed comprehension of the trends and patterns within the dataset by creating a thorough portrayal of the insights of the data. The chapter concludes in the use of data modelling approaches, including topic modelling via Latent Dirichlet Allocation (LDA) and sentiment analysis via the VADER algorithm. These efforts open the door for significant insights and discoveries in the analysis' later stages.

|  |  |
| --- | --- |
| **Chapter 4** | **Research Findings** |

# | Research Findings

## 4.1 Introduction

In-depth analyses of three different phone models from five separate phone brands—Apple, Samsung, OnePlus, Xiaomi and Realme are covered in this section

This section consists of:

* Amazon product reviews are classified into product and service reviews
* Sentiment behind the product and service review are analysed using VADER algorithm.
* A parallel sentiment review system is introduced along with the amazon rating.
* Comprehensive analysis of topic modelling using the Latent Dirichlet Allocation (LDA) method.

## 4.2 Classification of amazon reviews to product and Service reviews

As we discussed in the chapter 3, section 3.6 customers' opinions of the actual products and their experiences with the buying process are separated by Amazon's classification of reviews into "Product Reviews" and "Service Reviews." "Product Reviews" focus on the features, level of excellence, and functionality of the products, providing information on customer satisfaction and usefulness. However, "Service Reviews" cover topics like delivery time, packaging, and customer service, providing insight into the entire buying experience. This classification helps to understand how customers feel about the platforms and the products, offering useful information for product creation, improving customer support, and analysing market trends. This method includes the use of customized rules and keywords based on the Amazon reviews available in the dataset.

After the classification we obtained below results:

* Total Amazon reviews: 29,798
* Total product reviews: 28,374
* Total service reviews: 1424

As you can see in figure 16, based on the classification rules the reviews are labelled into product and service reviews. The labels are stored into a new Coolum named “Classification”. The labelling will help us understand the proportion of product and service reviews in the overall amazon reviews. The labelling will provide overview of the major service issues and product features related to amazon mobile brands.

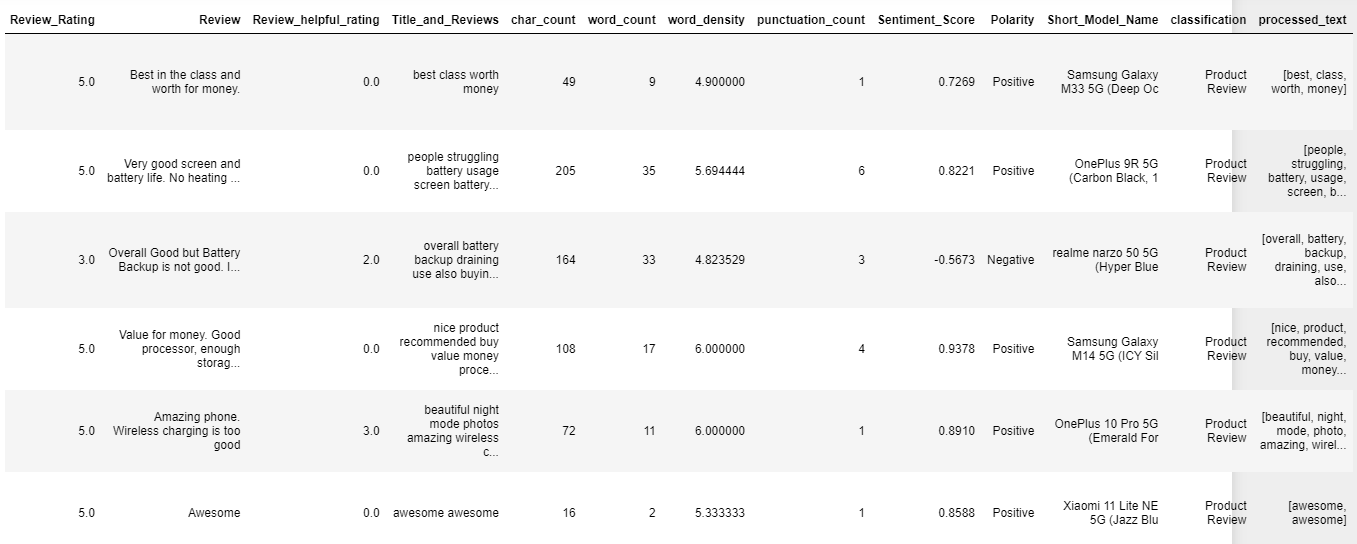


Figure 16,Product review and service review labelled data

The percentage of product and service reviews out of all Amazon reviews is depicted in the figure 17. It is clear from the figure that 95.2% of all reviews are for products. Only a small percentage of service reviews were written by customers. Despite their smaller proportion, service reviews are crucial for giving clients a better shopping experience. The service reviews will shed light on areas where the service needs to be enhanced as well as the necessary actions. Customers of Amazon who visit the website will also be aware of the problems experienced by prior users of Amazon products and services. The transparency of reviews will help the customer to take an informed purchase decision.

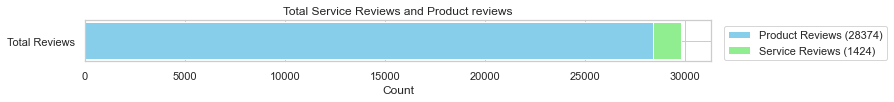


Figure 17,Overall product and service review

From the figure 18 it’s evident that in the dataset the large portion of the reviews is for the mobile brands such as Samsung, OnePlus and Xiaomi. The Xiaomi brand received a total of 10775 reviews with 10324 product reviews and 451 service reviews. While the Samsung brand received a total of 9766 reviews with 9233 product reviews and 533 service reviews. OnePlus brand received a total of 6937 reviews with 6663 product reviews and 274 service reviews. The Apple and realme brands collectively received a total of 2320 reviews with 2154 product reviews and 166 service reviews.

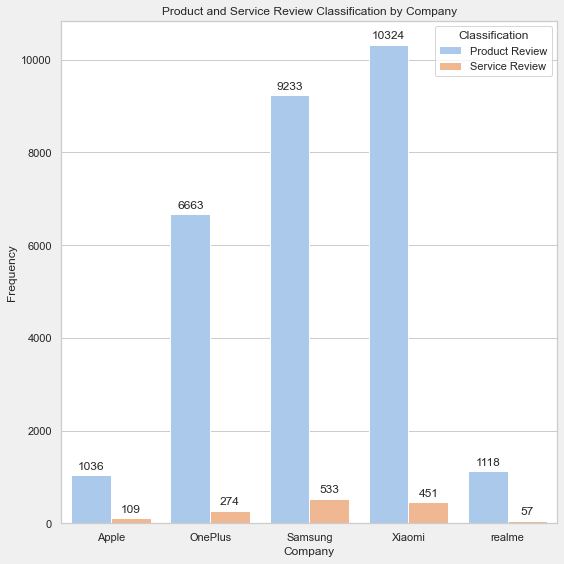


Figure 18, Product and service reviews for each model

We may use the labelled data in the section that follows to conduct an in-depth study of sentiment and commonly discussed topics. A more detailed understanding of the products and services provided by Amazon items may be provided by the labelled dataset. This study offers an insight of overall customer satisfaction and sentiment to help with product analysis, business strategy creation, and targeted changes.

## 4.3 Sentiment analysis using Vader algorithm

The reviews in column “Titles\_and\_Reviews” are cleaned and fed to Vader algorithm for finding the sentiment behind the text. In order to analyse sentiment in text, VADER particularly takes into account both the polarity (positive, negative, or neutral) and intensity of the sentiments represented. Finding the general sentiment patterns related to products, services, and customer experiences can be done with the help of VADER algorithm applied to Amazon reviews.

* Positive Sentiment Score: This score indicates the probability of positive sentiment in the text. It represents the proportion of words with positive sentiment to all words in the text.
* Negative Sentiment Score: Similar to the positive score, the negative score represents the likelihood of negative sentiment in the text.
* Neutral Sentiment Score: The neutral score measures the probability of neutral sentiment in the text. It calculates the proportion of words with neutral sentiment to all words.
* Compound Sentiment Score: The compound score provides an aggregated sentiment score that considers all positive, negative, and neutral sentiments. It ranges between -1 and 1. It’s a useful overall indicator of sentiment.

Figure 19 shows the newly created columns in the dataset The reviews are labelled as Positive, Negative or Neutral based on the sentiment score obtained from Vader. The Vader classified the total reviews into 19817 Positive reviews, 8198 Negative reviews and 1786 Neutral reviews. The labels are stored in column named “Polarity” and the sentiment score obtained is stored in column named “Sentiment score”.

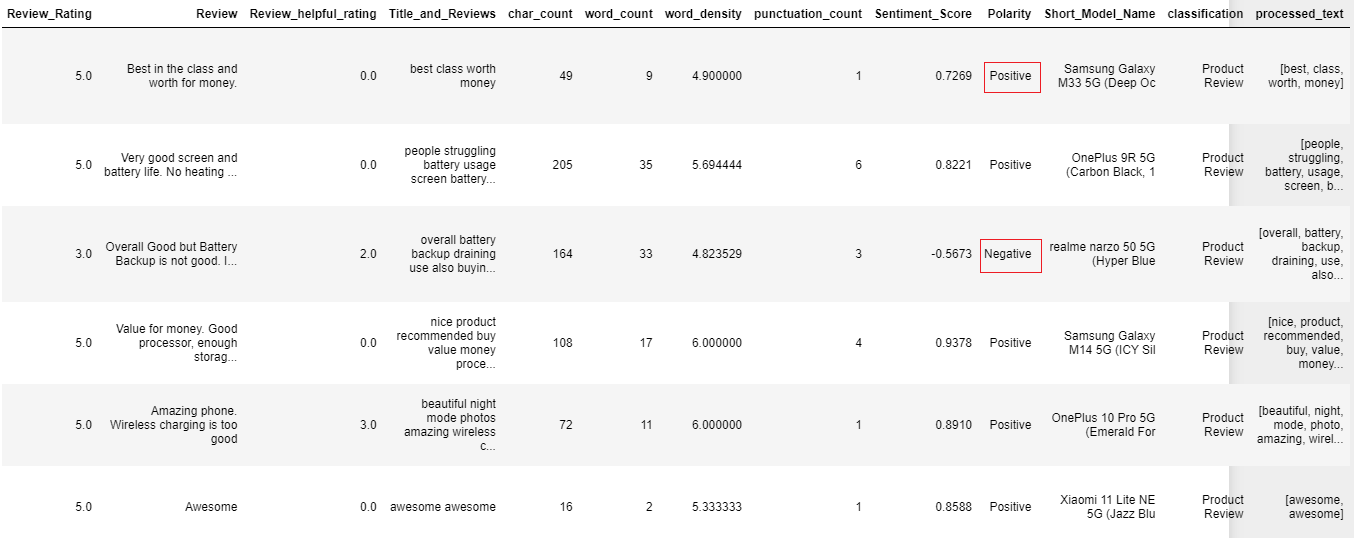


Figure 19,Dataset with product review and service review labels

From the Pie chart, figure 20, we can understand how the reviews are divided into Positive, Negative and Neutral reviews based on the compound sentiment score. The pie chart shows that from the overall reviews 66.5% of the reviews are biased to positive sentiment. At the same time, 27.5% of review data is related to Negative sentiment. From this figure and considering the dataset in hand, we can infer that most of the customers are satisfied with the products and services. Even though the positive sentiment is higher the negative aspects need to evaluated to provide better service to the customer. The later sections will explain more about the relation between the sentiment score and other columns in the dataset.

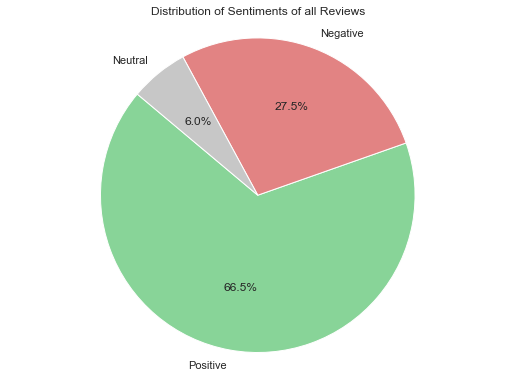


Figure 20,distribution of sentiments for overall reviews in the dataset

The sentiment score distribution for all of the reviews is shown in the figure 21. By analysing sentiment ratings and their frequency, one can gain an in-depth understanding of the customer's sentiments represented within a dataset. The sentiment score distribution shows the strength of each customer's opinion's feeling. It is evident that over 8000 reviews ranged in score from 0.05 to 1. The majority of the reviews are rated as positive. Negative reviews are those with a sentiment score between -0.05 and -1, which represent a smaller percentage of all reviews. The neutral comment shared a range of mixed emotions about the good or service, with a sentiment score of between -0.05 and 0.05.

This information makes it very evident that customers expressed a strongly positive opinion in their reviews. This demonstrates that majority of these customers was extremely satisfied with the products or services offered by Amazon.

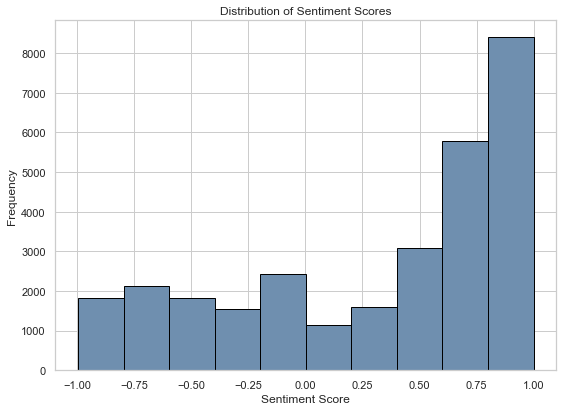


Figure 21,sentiment score distribution for all the reviews in the dataset

The Figure 22 visualize the relation between the average sentiment score and the review ratings. The x axis represents the review rating for each reviews axis represents the average sentiment score calculated by Vader algorithm. The visualization id created to understand the relation between the emotion behind the reviews and the ratings provided by the customers. The reviews 4 and 5 scored a sentiment value in range between 0 to 0.7. This indicates that, the customers were having a positive emotion while providing these ratings. The ratings 2 and 3 scored a sentiment score ranging in between 0.2 and -0.1. This value indicates a neutral sentiment. Some customers who rated their reviews using rating 4 and 5, express their opinions with sentiment score close to zero. which indicates that, even though the rating 5 is mostly used to express extreme satisfaction, customers used them to express their disagreements as well. This behaviour noticed in Customers want their experiences or opinions to be noticed by other users. This finding can be backed by analysing the sentiment score distribution of rating 1. The sentiment score for rating 1 ranges in between 0 to -0.9. This also implies that when a customer what other consumers to notice their opinions they tend to give rating which are in extreme ends. This human behaviour leads the customers to give ratings in 5 and 4 even though their reviews are not strongly positive.

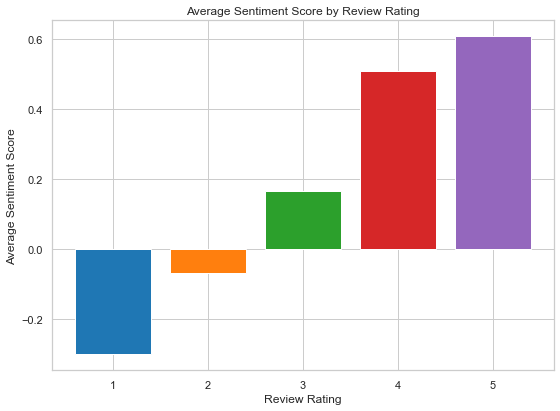


Figure 22,Averege sentiment score against the review rating

Figure 23 shows the sentiment of reviews for each smart phone models in the dataset. This visualization will be helpful to understand which model was more favourable reviews and which one got the most criticization We have 20 modes which are under brands such as Apple, Samsung, OnePlus, Xiaomi and Realme. From the visualization it is clear that Xiaomi, Samsung and OnePlus smart phone models received more negative reviews when compared to Apple and Realme smartphones. Xiaomi 13 pro, Samsung S23 and OnePlus 9R, Apple Iphone14 received a lot of positive response from the customers. These are all new phone in the market which got released after September 2022. They are in the market only for 6 months to the date the data is collected. The "Apple phone SE" received the most unfavourable reviews out of all the Apple handsets. The "OnePlus 10 Pro 5G" earned the most negative feedback of all the OnePlus smartphones. The "Samsung Galaxy M14" received the most negative comments of all the Samsung smartphones. The "Xiaomi 11 Lite" smartphone from Xiaomi earned numerous poor feedback.

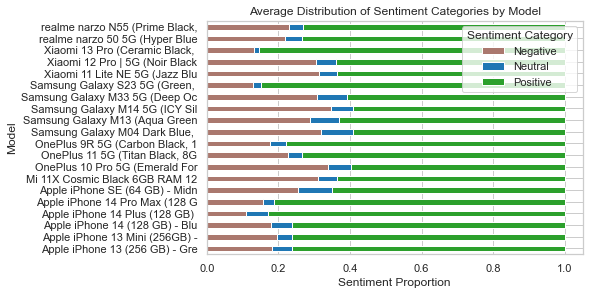


Figure 23,Sentiment proportion for each smartphone model in the dataset

In this figure 24 the visualization shows the sentiments behind products and amazon services. The X axis consist of sentiment categories and Y axis consist of frequency of each category. The data from “Polarity” and “Classification” columns were used to create this graph. The left diagram represents the product review sentiments and the right one represents the service review sentiment. From the left graph we can understand that, 18000 product review are biased to positive sentiments.7500 negative review was found among the product review. Neutral sentiment product reviews are less when compared to positive and negative reviews. From the right graph out of 1424 service review 700 reviews were found as positive. At the same time customers reported 620 negative service reviews. Even though the number of service issues reported as less, the cause of the issues needs to identified. By resolving the service issues, we can increase the customer satisfaction and attract more new customers to the platform.

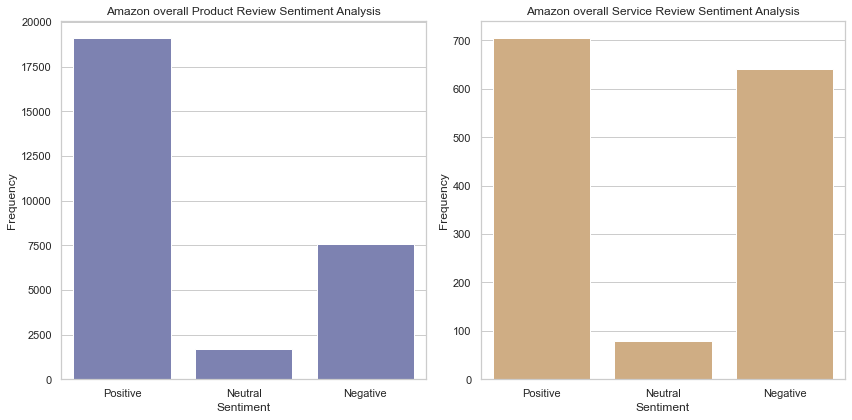


Figure 24,Sentiment distribution based on the product reviews and service reviews available in the review text from the whole dataset

**Corelation matrix heatmap**

In the figure 25, the correlation matrix of the dataset is represented. For the analysis numerical columns including extracted features are selected. The correlation value, often known as the correlation coefficient, is a statistical metric that expresses how linearly related two variables in a dataset are to one another. It shows how closely these variables' values vary together. The correlation coefficient, which runs from -1 to 1, sheds light on the nature and magnitude of the relationship.

* Positive Correlation (0 to +1): When the correlation between two variables is positive, it means that as one increases, the other increases as well. The two variables move in the same direction when the correlation value is closer to +1, indicating a strong positive link.
* Negative Correlation (0 to -1): A negative correlation value denotes a tendency for one variable to decrease as the other increases. When the two variables move in the opposite directions, there is a significant negative association, as indicated by a correlation value that is closer to -1.
* No Correlation (zero): The absence of a linear relationship between the variables is indicated by a correlation value of 0. Changes in one variable do not predict or have an impact on changes in the other.

While analysing the heatmap it is evident that the sentiment score has a positive correlation of 0.60 with review rating column. The review rating is the numerical expression of sentiment, while the review text is used to express the opinions in words. If the review rating is increasing the review rating also increases, which shows a positive correlation between the columns. This means the sentiment score found through Vader algorithm is correctly predicted the sentiments behind the review.

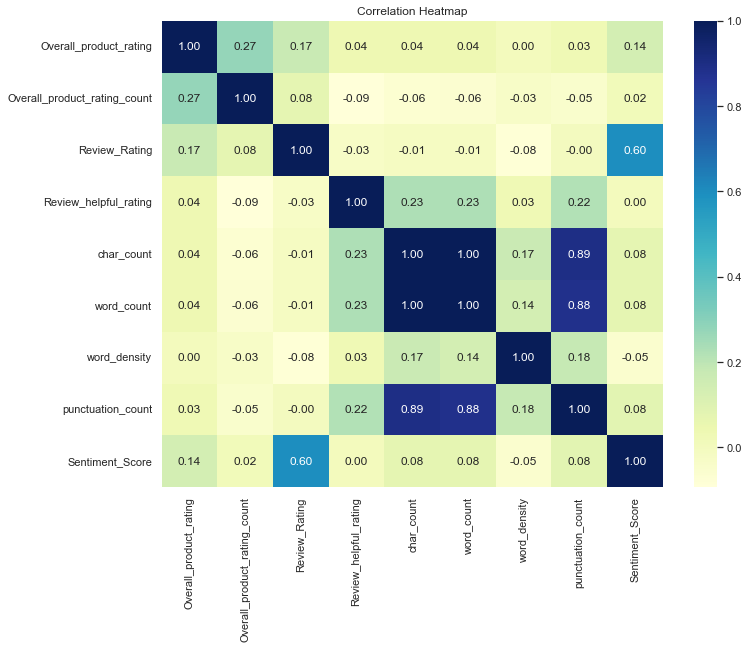


Figure 25, Corelation matrix Heatmap

## 4.4 Introducing a parallel sentiment rating for amazon

The main objective of this project is to create a parallel rating system for amazon products. To perform that, three smart phones models with higher number of reviews in the dataset is selected. Through this analysis we will be able understand the most discussed subjects in product and service review for an individua model.

The selected smartphone models are:

1. Xiaomi 11 Lite NE 5G (Jazz Blue 6GB RAM 128 GB Storage)
2. Samsung Galaxy M33 5G (Deep Ocean Blue, 8GB, 128GB Storage)
3. OnePlus 9R 5G (Carbon Black, 12GB RAM, 256 GB Storage)

### 4.4.1 Analysis for Xiaomi 11 Lite NE 5G (Jazz Blue 6GB RAM 128 GB Storage)

Xiaomi 11 Lite NE 5G is a smartphone created by Xiaomi company which s release on 15 September 2021 to the market. This product has a review rating of 4 stars from 25835 reviews in amazon website. In the 4819 reviews received in amazon, 4627 reviews are categorized into product reviews and 192 reviews are service reviews (Figure 26).

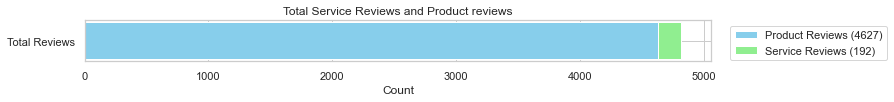


Figure 26,Total product and service reviews for Xiaomi 11 Lite NE 5G

The figure 27, shows the product and service setiment behind the Xiaomi 11 Lite NE 5G model. The left graph shows the product sentiment review with sentiment category in X axis and frequency in Y axis. From this graph it is evident that 2964 reviews are related to positive sentiment, while 1427 review are negative reviews. The right graph shows the service sentiment review with sentiment category in X axis and frequency in Y axis. From this graph it is evident that 102 reviews are related to positive sentiment, 82 review are negative reviews.

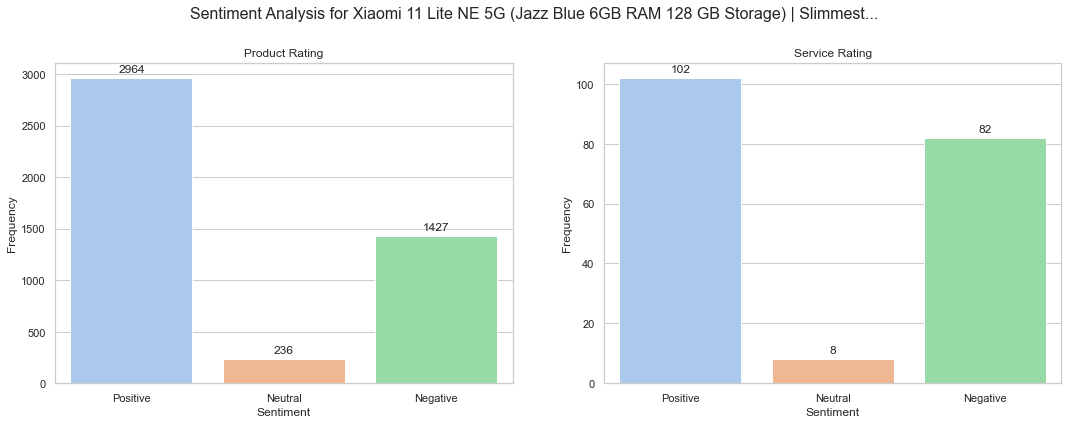


Figure 27,Product review and service review sentiment analysis for the Xiaomi 11 Lite NE 5G

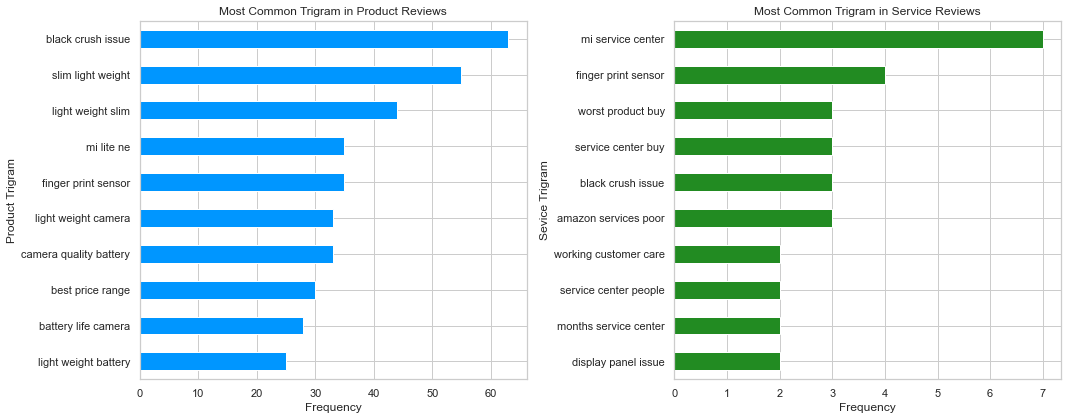
The figure 28, represents the Trigram of the product and service-related reviews. The trigram helps us to understand the most discussed words and its context. The most discussed words related to product review for Xiaomi 11 Lite NE 5G is, light weight, fingerprint sensor, camera, battery life and price rage. The most discussed service-related words for Xiaomi 11 Lite NE 5G are, fingerprint sensor, amazon service, display panel issues and issues related to service centre.

Figure 28,Most frequently discussed words for the Xiaomi 11 Lite NE 5G

### 4.4.2 Analysis for Samsung Galaxy M33 5G (Deep Ocean Blue, 8GB, 128GB Storage)

Samsung Galaxy M33 5G is a smartphone created by Samsung company which s release on 8 Aprill 2022 to the market. This product has a review rating of 4.1 stars from 25835 reviews in amazon website. In the 4869 reviews received in amazon, 4639 reviews are categorized into product reviews and 230 reviews are service reviews (Figure 29).

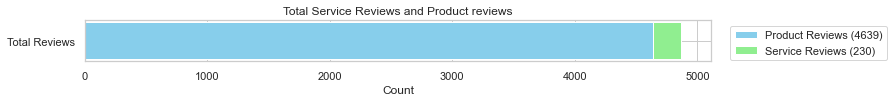
****

Figure 29,Total product and service reviews for Samsung Galaxy M33 5G

The figure 30 shows the product and service setiment behind the Samsung Galaxy M33 5G model. The left graph shows the product sentiment review with sentiment category in X axis and frequency in Y axis. From this graph it is evident that 2869 reviews are related to positive sentiment, while 1387 review are negative reviews. The right graph shows the service sentiment review with sentiment category in X axis and frequency in Y axis. From this graph it is evident that 103 reviews are related to positive sentiment, 114 review are negative reviews. The service review has a higher rate of negative reviews.

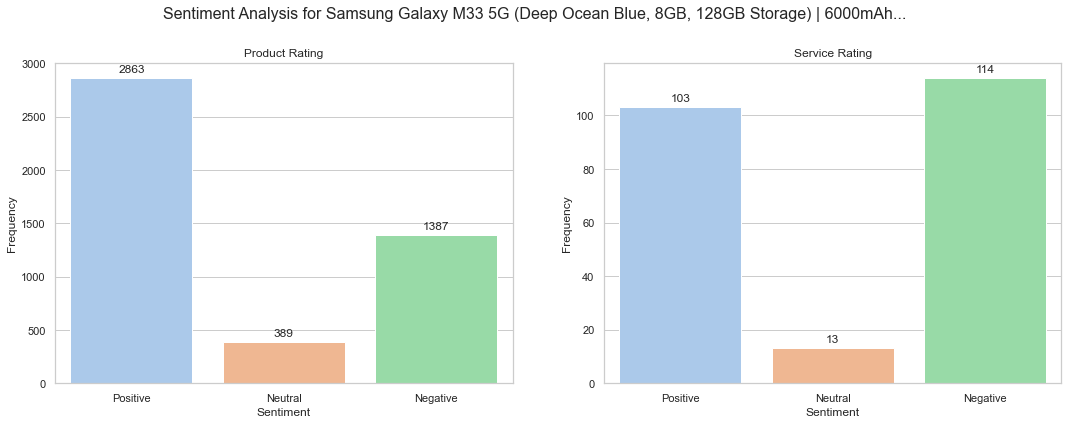


Figure 30,Product review and service review sentiment analysis for the Samsung Galaxy M33 5G

The figure 31 represents the Trigram of the product and service-related reviews. The trigram helps us to understand the most discussed words and its context. The most discussed words related to product review for Samsung Galaxy M33 5G model is ‘camera quality bad’, ‘battery life’, ‘battery drain’, and ‘value for money’. The most discussed service-related words for Samsung Galaxy M33 5G are, ‘Amazon service bad’, ‘worst service amazon’, ‘replacement available’, and ‘bad experience amazon’.

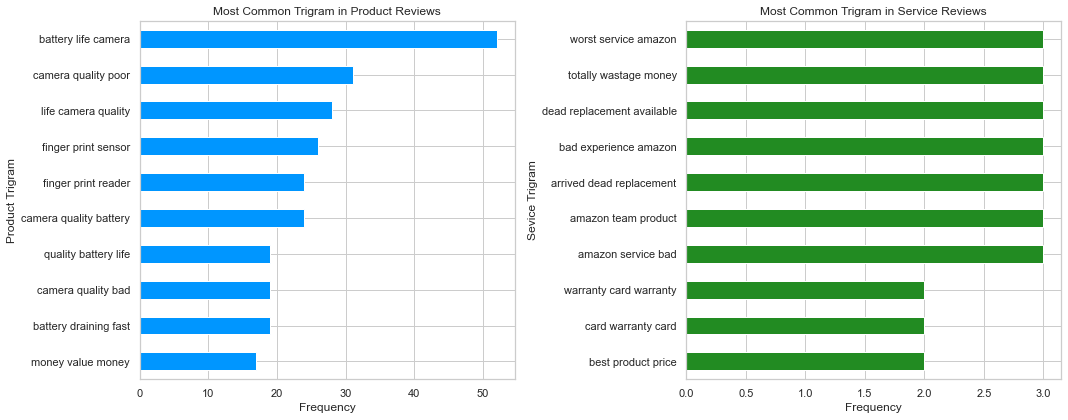
****

Figure 31,Most frequently discussed words for the Samsung Galaxy M33 5G

### 4.4.3 Analysis for OnePlus 9R 5G (Carbon Black, 12GB RAM, 256 GB Storage)

OnePlus 9R 5Gis a smartphone created by OnePlus company which s release on 15 September 2021 to the market. This product has a review rating of 4 stars from 25835 reviews in amazon website. In the 4819 reviews received in amazon, 4627 reviews are categorized into product reviews and 192 reviews are service reviews (Figure 32).

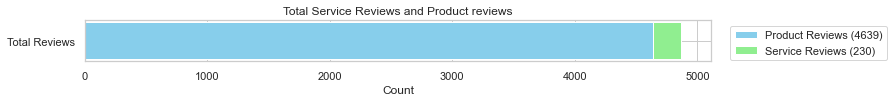


Figure 32,Total product and service reviews for OnePlus 9R 5G

The figure 33 shows the product and service setiment behind the OnePlus 9R 5G model. The left graph shows the product sentiment review with sentiment category in X axis and frequency in Y axis. From this graph it is evident that 3690 reviews are related to positive sentiment, while 803 review are negative reviews. There are 215 neutral product reviews as well. The right graph shows the service sentiment review with sentiment category in X axis and frequency in Y axis. From this graph it is evident that 90 reviews are related to positive sentiment, 50 review are negative reviews. There are 7 neutral reviews a well. The customers experienced a satisfactory feeling from the usage of the smartphone and service provided by amazon.

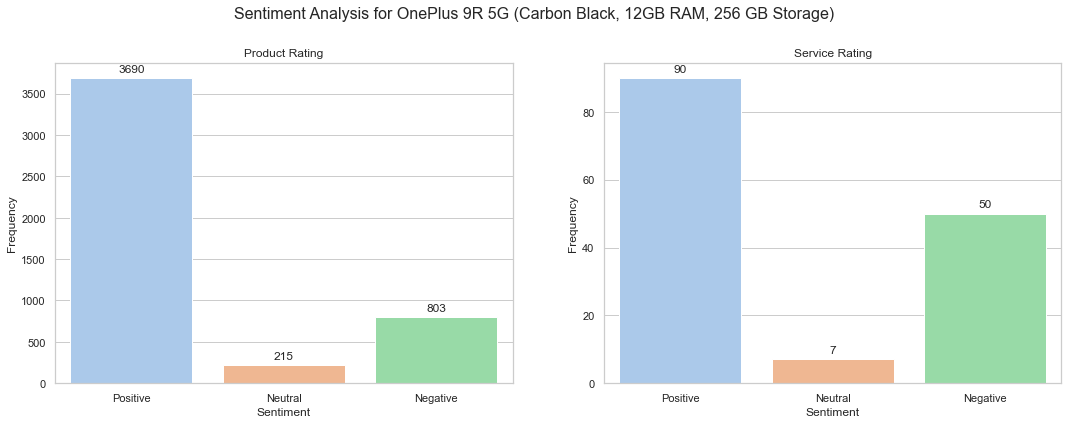


Figure 33,Product review and service review sentiment analysis for the OnePlus 9R 5G

The figure 34 represents the Trigram of the product and service-related reviews. The trigram helps us to understand the most discussed words and its context. The most discussed words related to product review for OnePlus 9R 5G are ‘battery drain issues’, ‘Fingerprint sensor’, ‘camera quality’ and ‘best price range’. The most discussed service-related words for OnePlus 9R 5G are, ‘Delivery time and packing’, and ‘Service centre technician’. Customers experienced a lot of issues from the One plus service centre and its technician.

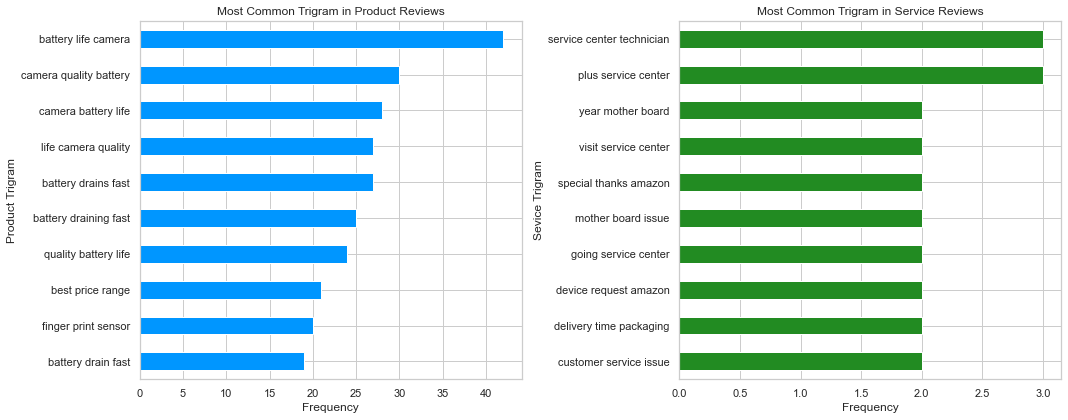


Figure 34,Most frequently discussed words for the OnePlus 9R 5G

## 4.5 Topic modelling

A statistical modelling technique called topic modelling is used to identify more or less abstract subjects in a set of documents. It uses machine learning to analyse text data automatically in order to identify cluster terms for a collection of texts. In order to uncover latent semantic structures in textual data, topic modelling is particularly popular in text mining. This unsupervised machine learning technique is appropriate for data exploration. In this section we will be analysing the results obtained from the topic modelling of the review dataset.

The section involves:

1. Tuning the topic model (LDA) to achieve maximum accuracy
2. Analysing Topic Bubble Chart
3. Analysing pyLDAvis visualization graph

**Coherence Score:** Coherence score measures the interpretability and clarity of topics generated by the model. It calculates the degree of semantic similarity between words within a topic. Higher coherence scores indicate more coherent and meaningful topics. Commonly used coherence measures include UMass and C\_V coherence.

**Perplexity:** Perplexity is a measure of how well the model predicts unseen data. It evaluates how well the topics learned by the model generalize to new documents. Lower perplexity values indicate better predictive performance of the model.

**Number of Topics:** Determining the right number of topics is crucial. If the number is too low, the topics might be too broad and fail to capture finer nuances. If the number is too high, the topics might become overly specific or may contain noise.

### 4.6 Tuning the topic model (LDA) to achieve maximum accuracy

Latent Dirichlet Allocation (LDA) is a topic model that can be tuned to produce the best outcomes by modifying a number of parameters and hyperparameters to enhance the precision and calibre of the topics produced. Since LDA is a probabilistic model, it may be improved to produce more insightful and understandable concepts. The quantity of topics is the key tuning variable. To strike the correct balance between having too few generally applicable topics and too many very specific topics, you'll need to experiment with various values of K.

To find the optimal number of topics, coherence score of 5 topics were analysed. From the figure it is evident that topic 4 achieved a coherence score of 0. 5314.After topic 4 the value starts to decrease, resulted in low coherence score and accuracy. Thus, the optimal K-Value is found to be 4 (Number of topics). Below figure 35, plots a line graph based on the coherence score obtained for 5 topics.

Coherence Score for 1 Topics: 0.3798

Coherence Score for 2 Topics: 0.4846

Coherence Score for 3 Topics: 0.5017

Coherence Score for 4 Topics: 0.5314

Coherence Score for 5 Topics: 0.5140

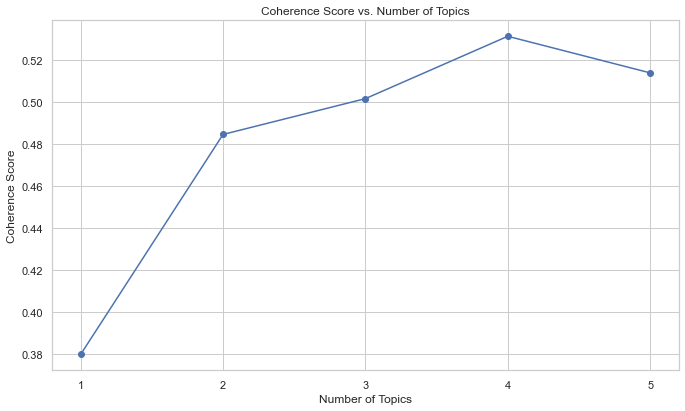


Figure 35,Line graph with coherence score for 5 topics

### 4.7 Analysing Topic Bubble Chart

As we see in the figure 36, bubble chart is a style of data visualization that represents data points in a two-dimensional space using circles (bubbles). The horizontal position, vertical position, and size of the bubble are the traditional three data dimensions for each bubble in the chart. The bubble's X and Y coordinates are determined by its horizontal and vertical positions, and its size is sometimes employed to represent a third quantitative variable.

Based on the K value of 4 the LDA model achieved a coherence core of 0.53 and a perplexity score of -6.67. That means the topics generated by the model has more meaningful insights and better predictive performance.

Perplexity: -6.6773733082280025

Coherence Score: 0.5381815425369166

**Bubble Size:**

Larger bubbles indicate higher values of the third variable being represented. This could be a numerical measurement, frequency, or any other quantitative value. In this figure the bubbles for Topic 1 are bigger hen compared to other topics. That means the variables in the topic 1 is more in numbers when compared to variables in other topics.

**Correlation and Patterns:**

Bubbles that are close together might indicate positive correlation or similarity between the variables. Bubbles that are far apart might indicate a lack of correlation or distinct differences. Here all the bubbles are distant to each other. That means the correlation of the topics between them is distinct.

**Position of bubbles in quadrant:**

1. Quadrant I (Positive-Positive): Bubbles located in this quadrant have positive values on both the X and Y axes. These bubbles represent data points where both variables are positively correlated or have positive impacts. There are no bubbles in quadrant 1 that means there are variable than are positively correlated to each other.
2. Quadrant II (Negative-Positive): Bubbles in this quadrant have negative values on the X-axis (left side) and positive values on the Y-axis (upper side). These bubbles represent scenarios where one variable has a negative impact (hence the negative X value), but the other variable has a positive impact. Variables in Topic 3 is located in quadrant 2, which means the Y variable values increase when the x variable values increase.
3. Quadrant III (Negative-Negative): Bubbles in this quadrant have negative values on both the X and Y axes. These bubbles represent data points where both variables have negative correlations or negative impacts. Variables in Topic 2 and 1 is located in quadrant 3, two bubbles are in the same quadrant of a scatter plot or bubble chart, it indicates that both data points share similar relationships between the variables being plotted. It suggests a negative correlation or negative relationship between the variables for both data points.
4. Quadrant IV (Positive-Negative): Bubbles in this quadrant have positive values on the X-axis (right side) and negative values on the Y-axis (lower side). These bubbles indicate scenarios where one variable has a positive impact, but the other variable has a negative impact. Variables in Topic 4 is located in quadrant 4, which means the X variable values increase when the Y variable values increase.

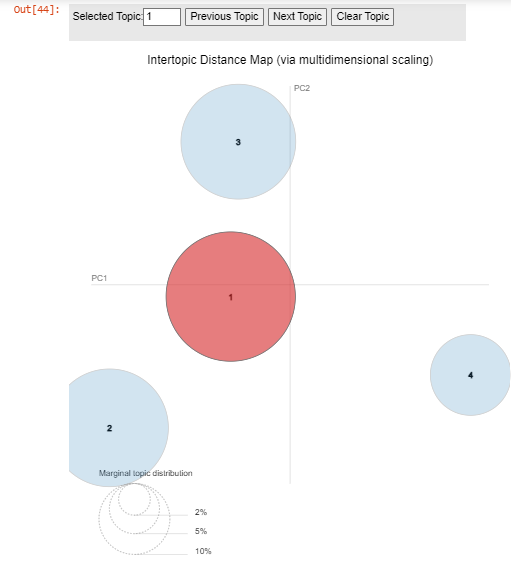


Figure 36,LDA Bubble chart

### 4.8 Analysing pyLDAvis visualization graph

This chart displays the distribution of topics in the corpus. Each bar represents a topic, and its height corresponds to the prevalence of that topic across all documents. The x-axis represents the topics, while the y-axis represents the proportion of the topic in the corpus. You can use this chart to identify dominant topics and understand their significance in the dataset. Here for the analysis, we took the whole amazon review data. This is done to understand the common topics discussed by customers on all products and services

**Topic 1: User experience**

In topic 1, most of the discussed tokens are related to user experiences. The Figure 37 can be interpreted based on the most occurred words.33.3% of all the tokens are available in topic 1. The tokens like experience, user, dip lay, features, using, day, better, feel suggested the topic is more related to user experiences with the product sold. Companies can use this data to identify the different experiences of users. By understanding in which areas, the customers mainly focused on user experience can we used as a selling point.

**Topic 2: Product features**

In topic 2 the customers mostly discussed about the camera quality, display speed of the product, charging speed, performance of the phone, battery and camera. This indicates that the topic mainly discusses about the product features.27.5% of the tokens are available in topic 2. The most discussed features can used to understand what the customer required in the future products. In which improvement need to done.

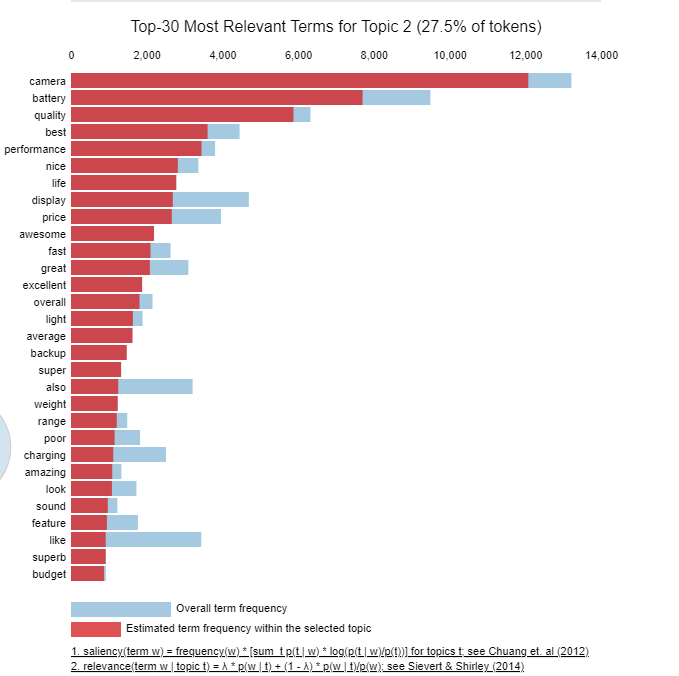
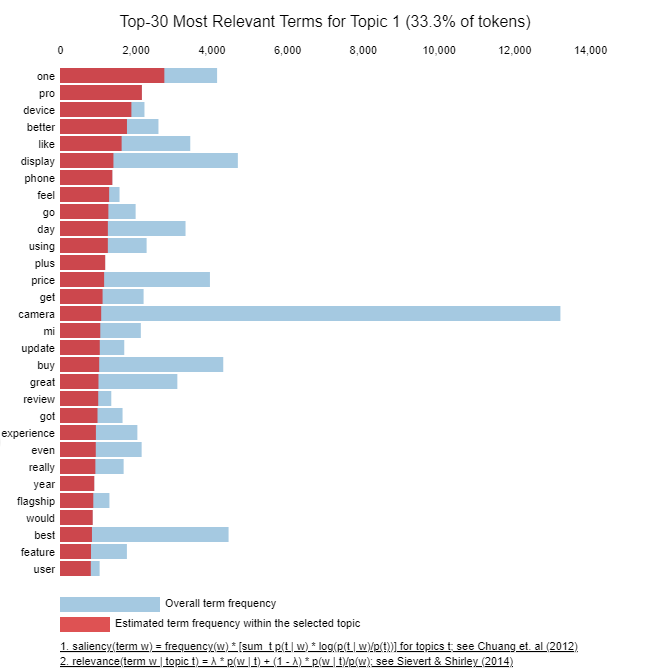


Figure 37,LDA most discussed topics in Topic and Topic 2

**Topic 3: Product issues**

In figure 38, topic 3 the customers mostly discussed about the heating issues, phone hanging, sensor, screen working and so on. This indicates that the topic mainly discusses about the product issues.26.52% of the tokens are available in topic 3 The most discussed product issues can used to understand and resolve them for future customers.

**Topic 4: Service issues**

In figure 38 topic 4 the customers mostly discussed about the amazon customer service, exchange, delivery, service, replacement and return. This indicates that the topic mainly discusses about the service issues.13% of the tokens are available in topic 4 The most discussed service issues can used to understand and resolve them for future customers in the amazon platform.

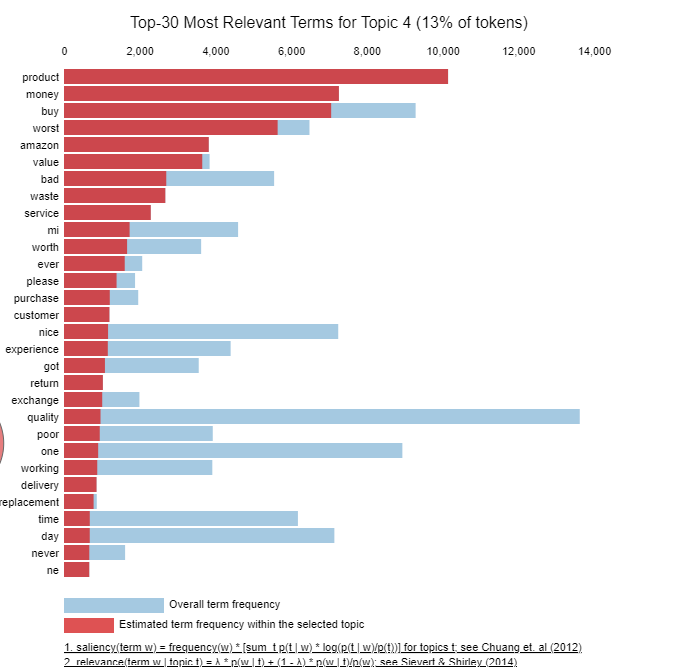
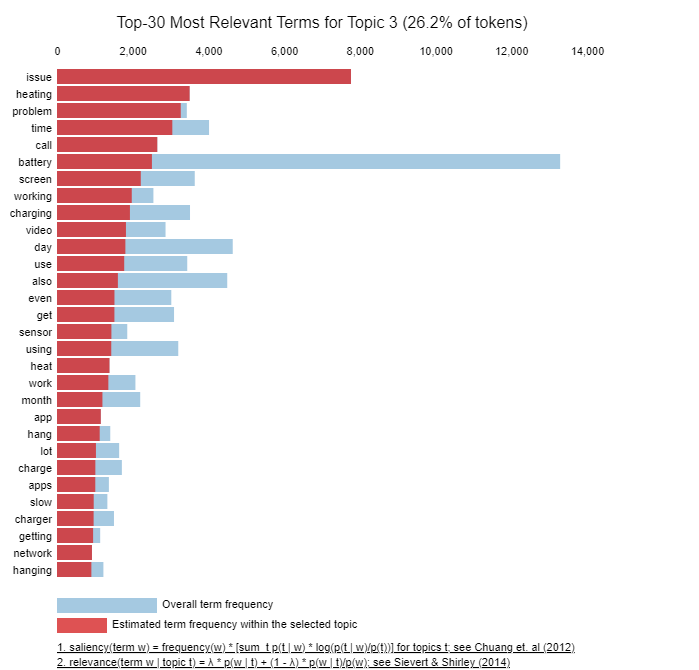


Figure 38, LDA most discussed topics in Topic 3 and Topic 4

## 4.9 Summary

In this chapter the study mainly focused on the sentiment analysis and topic modelling approaches using Latent Dirichlet Allocation (LDA). To gain deeper insights from the gathered data, the research goes into the use of two separate methods, VADER and LDA. The raw data is first presented in the chapter, after which it is thoroughly examined utilizing these two methods.

The reviews are categorized and labelled into service and product reviews. Later the categorized data is fed into Vader algorithm. From the analysis it is visible that about ,66.5% of the total reviews are positive sentiment reviews and 27.5% of the review are negative reviews. To determine the emotional content of the data, VADER, a sentiment analysis tool, is used. The sentiment analysis's findings highlight common attitudes, such as positivity, negativity, or neutrality, within the data points. The research obtains a quantitative comprehension of the emotional undercurrents in the dataset by measuring these sentiments.

In the later section of the chapter a parallel review system based on sentiment analysis is introduced for Samsung, Xiaomi and OnePlus models. To create rating system using sentiment analysis, trigrams are used to identify the most discussed words in the reviews. Along with the distribution of sentiment labelled for products and service reviews for each model is evaluated. The use of Latent Dirichlet Allocation (LDA) as a topic modelling technique is also explored in this chapter. To find recurring themes and subjects in the data, LDA is utilized. The study identifies latent patterns that might not be immediately obvious using probabilistic modelling. These subjects provide insightful information on the underlying themes and ideas that underlie the dataset.

This chapter sheds light to the insights generated through the sentiment analysis using Vader algorithm. Topic modelling provided detailed analysis on the most discussed topics and content in the review texts.

|  |  |
| --- | --- |
| **Chapter 5** | **Conclusions and Recommendations** |

# | Conclusions and Recommendations

## 5.1 Conclusion

The study was focused on extracting the amazon reviews from the platform to analyse the sentiment behind them. The main aim of this work is to understand the major topics discussed by the customers in the reviews. Thus, by evaluating the sentiment and major topics behind the reviews, we can improve the quality of the products and service provided to customers. This will improve the customer traffic to the amazon platform, which will eventually increase the credibility of the products sold.

## 5.2 Answers derived to research questions

Potential customers' future purchasing decisions are significantly influenced by past customer feedback. These reviews act as a kind of link between previous customer experiences and the decisions made by potential customers. Future clients may feel more secure and confident after reading a positive review that is packed full of compliments on the quality of the goods, the superiority of the service, and how satisfied they were overall. In contrast, negative feedback could raise concerns and discourage potential customers from making a purchase. Consumers' impressions, expectations, and ultimately decisions are shaped by the sum of these reviews, which creates a combined picture of the product or service in their minds.

In the amazon reviews customers usually provides the feedback related products they brought as well as the services. This insight was the basic foundation of this thesis. Through analysis of amazon, and though carefully formulated rules reviews were categorized into Product reviews and service reviews. Analysis of the sentiment gap between these two categories can provide priceless information. Positive reviews of a product may reflect a customer's satisfaction with its usefulness and quality. A lower rating in service reviews, however, can point to areas where delivery or customer service procedures need to be improved. On the other hand, a spike in positive customer experiences may signify great service, while a decline may draw attention to potential service-related issues.

These categorized reviews were cleaned using several data preprocessing techniques providing a cleaned and transformed data. This processed review text then fed into VADER for sentiment analysis to find the sentiment score of each review text. This sentiment score is analysed to form labels such as positive, negative and natural to categorize the text based on polarity. Topic modelling using Latent Dirichlet Allocation model is used to find the most discussed topics in the product and service reviews. Several studies were conducted using amazon reviews, to find the sentiments behind the review text using several methods like Textblob, Vader, AFFIN score and so on. A thorough investigation was done on the different approaches used in this subject. Based on the analysis, it is evident that no one attempted to combine the categorization of amazon review into product and service along with VADER and LDA modelling technique. The VADER (Valence Aware Dictionary and Sentiment Reasoner) algorithm has proven to be incredibly successful at identifying sentiment in Amazon customer reviews of goods and services. VADER, which was created especially for social media writing, is excellent at capturing sentiments in brief, informal, and context-rich texts—features that are frequently present in online reviews. The use of topic modelling on Amazon reviews reveals underlying semantic structures as topics, each of which represents a particular subject that is common throughout the dataset. These subjects are made up of a list of keywords and their corresponding probabilities, which show how likely it is that a word will appear in the topic. The discovery of common issues, trends, and attitudes that can be hidden by the amount and diversity of reviews is made possible by the extraction of the underlying topics.

Sentiment analysis and topic modelling can be effective tools for boosting customer interaction, optimizing conversion rates, and fine-tuning targeted advertising campaigns on Amazon. The feelings of customer reviews related to certain products can be evaluated using sentiment analysis. This knowledge can help advertisers craft their ads such that they emotionally connect with consumers. Customer reviews' underlying topics are found via topic modelling. Based on these themes, advertisers can divide their target market and produce targeted ads that are relevant to their targeted customer's preferences and interests.

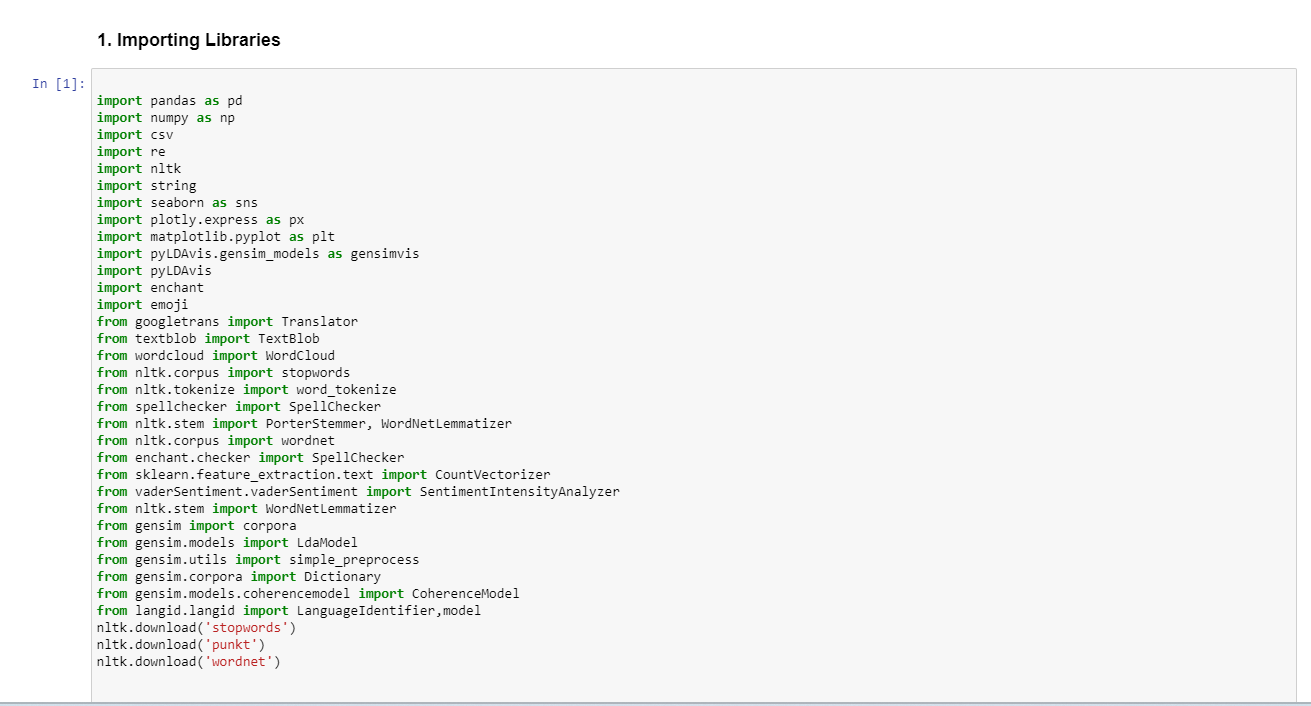
## 5.3 Limitations

The study is able to analyse the data and identify valuable insights from amazon review text. Even though the study was a success, there are some limitations to the research. The results found can be sample bias, because the research is based on a particular dataset of Amazon reviews, which could not accurately reflect the full range of customer thoughts and opinions. The results' generalizability may be affected by variations in review authenticity, quality, and diversity. Vader algorithm was able to identify the sentiment behind the review text successfully for almost 90% of the time. In some scenarios Vader faced difficulty to understand the nuances of irony, sarcasm, and dependent on context feelings which led to incorrect sentiment labelling. The rules formulated to categorize the product and service reviews also faced some issues in labelling the review text correctly. The diversity of the review text made it difficult to create a generalized rule which will work for all the reviews available in the dataset.

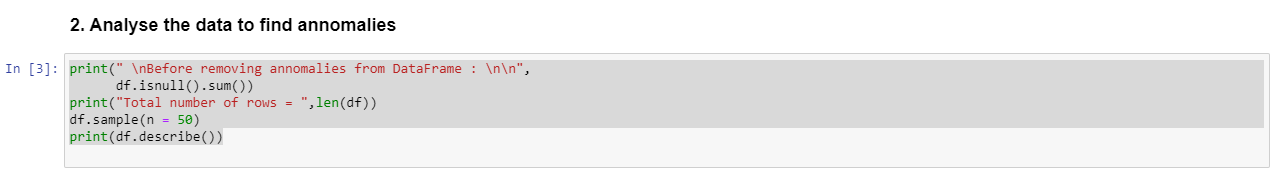
## 5.4 Future work

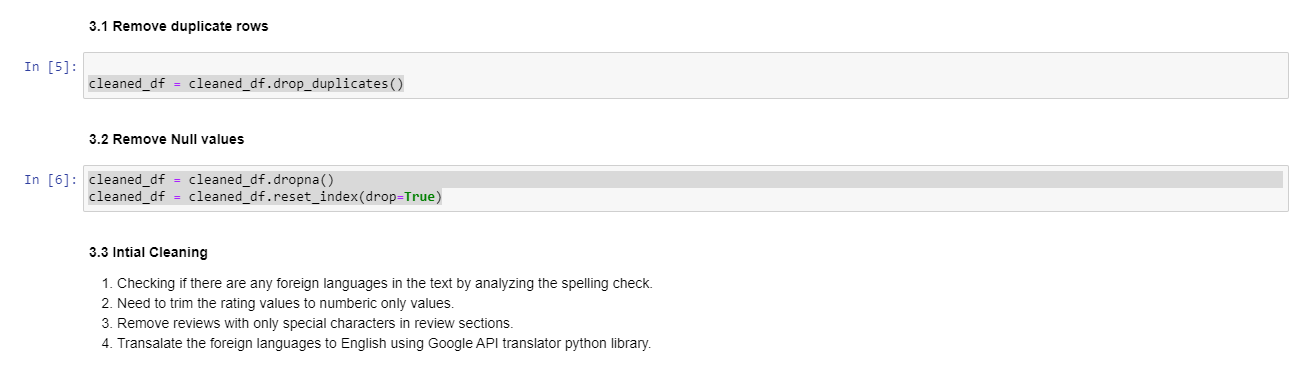
The future works involves in improving the accuracy of sentiment analysis and the categorization of amazon reviews into product and service reviews. Through this analysis we acquired enough data which is labelled using Vader algorithm. This labelled data can be used for machine leaning purpose to train and generate labels for the new amazon reviews for accurately. Since Naive bayes classifier is excellent for binary classification it can be used to classify the review text more accurately. The same method can be used for Product and service review classification to improve the labelling accuracy for future analysis.

# Appendix 1



















# References

|  |
| --- |
| Anna Baluch, 2023. 38 E-Commerce Statistics Of 2023. [Online] |
| Available at: <https://www.forbes.com/advisor/business/ecommerce-statistics/>  [Accessed 02 08 2023]. |
| Bickart, B. and Schindler, R.M., 2001. Internet forums as influential sources of consumer information. Journal of interactive marketing, 15(3), pp.31-40. |
| Chevalier & Mayzlin, 2006. The effect of word of mouth on sales: online book reviews. Sage Journals, 43(3), pp. 345-354. |
| Zan Mo, Y.-F. L. P. F., 2015. Effect of online reviews on consumer purchase behavior. Journal of Service Science and Management, 8(03), p. 419.  Zhang, K.Z., Cheung, C.M. and Lee, M.K., 2014. Examining the moderating effect of inconsistent reviews and its gender differences on consumers’ online shopping decision. International Journal of Information Management, 34(2), pp.89-98. |

Maheswaran, D. and Meyers-Levy, J., 1990. The influence of message framing and issue involvement. Journal of Marketing research, 27(3), pp.361-367.

Bautin, M., Vijayarenu, L. and Skiena, S., 2008. International sentiment analysis for news and blogs. In Proceedings of the International AAAI Conference on Web and social media (Vol. 2, No. 1, pp. 19-26).

Bollen, J., Mao, H. and Zeng, X., 2011. Twitter mood predicts the stock market. Journal of computational science, 2(1), pp.1-8.

Xu, T., Peng, Q. and Cheng, Y., 2012. Identifying the semantic orientation of terms using S-HAL for sentiment analysis. Knowledge-Based Systems, 35, pp.279-289.

Ding, X., Liu, B. and Yu, P.S., 2008, February. A holistic lexicon-based approach to opinion mining. In Proceedings of the 2008 international conference on web search and data mining (pp. 231-240).

Ray, P. and Chakrabarti, A., 2022. A mixed approach of deep learning method and rule-based method to improve aspect level sentiment analysis. Applied Computing and Informatics, 18(1/2), pp.163-178.

Hutto C, Gilbert E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. InProceedings of the international AAAI conference on web and social media 2014 May 16 (Vol. 8, No. 1, pp. 216-225).

Copaceanu, A.M., 2021. Sentiment Analysis Using Machine Learning Approach. Ovidius University Annals, Economic Sciences Series, 21(1), pp.261-270.

Alharbi, N.M., Alghamdi, N.S., Alkhammash, E.H. and Al Amri, J.F., 2021. Evaluation of sentiment analysis via word embedding and RNN variants for Amazon online reviews. Mathematical Problems in Engineering, 2021, pp.1-10.

Hasan, A., Moin, S., Karim, A. and Shamshirband, S., 2018. Machine learning-based sentiment analysis for twitter accounts. Mathematical and computational applications, 23(1), p.11

Naf'an, M.Z., Bimantara, A.A., Larasati, A., Risondang, E.M. and Nugraha, N.A.S., 2019. Sentiment analysis of cyberbullying on instagram user comments. Journal of Data Science and Its Applications, 2(1), pp.38-48.

Haque, T.U., Saber, N.N. and Shah, F.M., 2018, May. Sentiment analysis on large scale Amazon product reviews. In 2018 IEEE international conference on innovative research and development (ICIRD) (pp. 1-6). IEEE.

Bhatt, A., Patel, A., Chheda, H. and Gawande, K., 2015. Amazon review classification and sentiment analysis. International Journal of Computer Science and Information Technologies, 6(6), pp.5107-5110.

Pandey, P. and Soni, N., 2019, February. Sentiment analysis on customer feedback data: Amazon product reviews. In 2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon) (pp. 320-322). IEEE.

Kumar, S., Yadava, M. and Roy, P.P., 2019. Fusion of EEG response and sentiment analysis of products review to predict customer satisfaction. Information Fusion, 52, pp.41-52.

Kausar, S., Huahu, X.U., Ahmad, W. and Shabir, M.Y., 2019. A sentiment polarity categorization technique for online product reviews. IEEE Access, 8, pp.3594-3605.

Thakkar A, Mungra D, Agrawal A, Chaudhari K. Improving the performance of sentiment analysis using enhanced preprocessing technique and Artificial Neural Network. IEEE transactions on affective computing. 2022 Sep 15;13(4):1771-82.

Pradha, S., Halgamuge, M.N. and Vinh, N.T.Q., 2019, October. Effective text data preprocessing technique for sentiment analysis in social media data. In 2019 11th international conference on knowledge and systems engineering (KSE) (pp. 1-8). IEEE.

Chaithra, V.D., 2019. Hybrid approach: naive bayes and sentiment VADER for analyzing sentiment of mobile unboxing video comments. International Journal of Electrical and Computer Engineering (IJECE), 9(5), pp.4452-4459.

Nguyen, H., Veluchamy, A., Diop, M. and Iqbal, R., 2018. Comparative study of sentiment analysis with product reviews using machine learning and lexicon-based approaches. SMU Data Science Review, 1(4), p.7.

Woodall, R., Kline, D., Modaresnezhad, M. and Vetter, R., 2021. A cloud-based system for scraping data from amazon product reviews at scale. In Proceedings of the Conference on Information Systems Applied Research, Washington DC, USA.

Han, S. and Anderson, C.K., 2021. Web scraping for hospitality research: Overview, opportunities, and implications. Cornell Hospitality Quarterly, 62(1), pp.89-104.

Huang, Y.F. and Li, Y.H., 2019, December. Sentiment translation model for expressing positive sentimental statements. In 2019 International Conference on Machine Learning and Data Engineering (iCMLDE) (pp. 79-84). IEEE.

Cohen, J., 2013. Statistical power analysis for the behavioral sciences. Academic press.

Chambers, J. M., Cleveland, W. S., Kleiner, B., & Tukey, P. A. (1983). Graphical Methods for Data Analysis. Wadsworth & Brooks/Cole.

Jurafsky, D., & Martin, J. H. (2019). Speech and Language Processing (3rd ed.). Pearson.

Nakayama, M. and Wan, Y., 2019. The cultural impact on social commerce: A sentiment analysis on Yelp ethnic restaurant reviews. *Information & Management*, *56*(2), pp.271-279.

Abah, J.O., 2021. *Sentiment Analysis of Amazon Electronic Product Reviews using Deep Learning* (Doctoral dissertation, Dublin Business School).

Hamdallah, A.R., 2021. Amazon Reviews using Sentiment Analysis.

Rani, S. and Kumar, M., 2021. Topic modeling and its applications in materials science and engineering. *Materials Today: Proceedings*, *45*, pp.5591-5596.

Liu, L., Tang, L., Dong, W., Yao, S. and Zhou, W., 2016. An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus*, *5*(1), pp.1-22.

Weston, S.J., Shryock, I., Light, R. and Fisher, P.A., 2023. Selecting the Number and Labels of Topics in Topic Modeling: A Tutorial. *Advances in Methods and Practices in Psychological Science*, *6*(2), p.25152459231160105.

Silva, C.C., Galster, M. and Gilson, F., 2021. Topic modeling in software engineering research. *Empirical Software Engineering*, *26*(6), p.120.

Vayansky, I. and Kumar, S.A., 2020. A review of topic modeling methods. *Information Systems*, *94*, p.101582.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

Arora, R., Ge, R., Halpern, Y., Mimno, D., Moitra, A., Sontag, D., & Wu, Y. (2013). A practical algorithm for topic modeling with provable guarantees. In Proceedings of the 30th International Conference on Machine Learning (ICML-13) (pp. 280-288).

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993-1022.

Guo, Y., Barnes, S. J., & Jia, Q. (2016). Mining Meaning from Online Ratings and Reviews: Tourist Satisfaction Analysis Using Latent Dirichlet Allocation. Tourism Management.

 Farkhod, A., Abdusalomov, A., Makhmudov, F., & Cho, Y. I. (2021). LDA-Based Topic Modeling Sentiment Analysis Using Topic/Document/Sentence (TDS) Model. Applied Sciences, 11(23), 11091.

 MDPI. (2021). LDA-Based Topic Modeling Sentiment Analysis Using Topic/Document/Sentence (TDS) Model. Retrieved from <https://www.mdpi.com/2076-3417/11/23/11091>

Zvornicanin, E., 2022. When coherence score is good or bad in topic modeling? Baeldung on Computer Science.