# 

# 

**TEXT ANALYSIS ON SMARTPHONE ONLINE REVIEWS**

**AZIEM HAZIQ BIN ABDUL MALEK**

**RAIHAN FARISS BIN RIZAL HELMY**

**MAHFUZ BIN YUSRI**

**BACHELOR OF SCIENCE (HONS.) STATISTICS**

**FACULTY OF COMPUTER AND MATHEMATICAL SCIENCES**

# ACKNOWLEDGEMENT

IN THE NAME OF ALLAH, THE MOST GRACIOUS, THE MOST MERCIFUL

First and foremost, we want to thank and praise Allah for showering His blessings on us throughout our study work, allowing me to complete our Final Year Project, ”Text Analysis on Smartphone Online Reviews”.

We would like to dedicate our research to our loving supervisor, Madam Noorezatty Binti Mohd Yusop, for all of her advice, assistance, support, encouragement, and understanding. She is also the person who has aided us the most when we have been unable to see the light at the end of the tunnel. Her support was quite helpful. She has shown us how to conduct research and present our findings in the clearest and most simple manner possible. Working and studying under her direction was a great honour and privilege.

Apart from our advisor, we'd like to express our gratitude to Madam Zaitul Anna Melisa binti Md Yasin, our professor for the Final Year Project subject, for her encouragement, support, and advice during the preparation and completion of this study. We would also like to thank her for her time, as she was always available for advice whenever we needed it.

Last but not least, we want to express our gratitude to our parents for their unwavering support in ensuring that we are able to complete our study in a physically, mentally, and emotionally healthy manner. We would also like to extend our appreciation to everyone who helped us finish our Final Year Project, whether directly or indirectly. We will be eternally grateful for all of your help and support.

# 

[**CHAPTER 1  
INTRODUCTION**](#_bctmm54ooas2) **4**

[1.1 BACKGROUND OF STUDY](#_ws01lww6ikce) 4

[1.2 PROBLEM STATEMENT](#_lnb267mptfr8) 6

[1.3 RESEARCH OBJECTIVES](#_gzyhdj4bzox2) 7

[1.4 RESEARCH QUESTIONS](#_rszrarkztym8) 7

[1.5 SCOPE AND LIMITATION OF STUDY](#_shsg55mz0ivw) 8

[1.6 SIGNIFICANCE OF STUDY](#_r5dd0q5af0k) 8

[**CHAPTER 2  
LITERATURE REVIEW**](#_ql8qkm5a0462) **9**

[2.1 INTRODUCTION](#_1h4imi71mu53) 9

[2.2 MOVING TO E-COMMERCE](#_fvjbdhoktbl4) 9

[2.3 THE ANALYSIS OF ONLINE CUSTOMER REVIEWS](#_5j2nwisr4b61) 10

[2.5 CONCLUSION](#_2xjw08ankrei) 12

[**CHAPTER 3  
METHODOLOGY**](#_mekty071djk4) **13**

[3.1 INTRODUCTION](#_9yyl01snmt80) 13

[3.2 SOURCE OF DATA](#_wm6tf9x4fykb) 13

[3.3 TEXT MINING ANALYSIS](#_pwl33240doo4) 13

[3.3.1 Web scraping process and tools](#_1bm7alrpvvk8) 14

[3.3.2 Cleaning text data](#_79y7ibv3kezz) 15

[3.3.2.1 Tokenization](#_6a1khbndm4y6) 15

[3.3.2.2 Filtering](#_5und8lorhsbu) 16

[3.3.2.3 Stop word removal](#_nzbus0qelgik) 16

[3.3.2.4 Stemming](#_szo4gyujlqzr) 16

[3.3.3 Term-document matrices](#_srcc4uo65e2t) 17

[3.3.4 Text mining analysis](#_n1qn1xk3jh88) 17

[3.3.4.1 Descriptive analysis](#_t1bw9dfh39d5) 17

[3.3.4.2 Sentiment analysis](#_1pes4ankr63y) 18

[3.3.4.3 Association of words](#_hsnbppche7ue) 19

[3.3.4.4 Clustering Analysis](#_6amtf42sev0y) 20

[3.4 SUMMARY OF DATA ANALYSIS](#_7es0ng4txzgq) 21

[**CHAPTER 4  
RESULT AND ANALYSIS**](#_kzh4qx7xgjmn) **23**

[4.1 INTRODUCTION](#_ua527l5vviwo) 23

[4.2 DESCRIPTIVE ANALYSIS](#_bg1x389yhc48) 23

[4.3 POSITIVE AND NEGATIVE FEEDBACKS](#_umkd96um5gsf) 28

[‘4.4 ASSOCIATION BETWEEN REVIEWS](#_qv94methtgv) 29

[**CHAPTER 5  
CONCLUSION AND RECOMMENDATION**](#_77zh9g7540tt) **30**

[5.1 CONCLUSION](#_a32bepyc4kxf) 30

[5.2 RECOMMENDATION](#_km8qkc1nukjt) 30

[REFERENCES](#_644m5ra6wph2) 32

# 

# 

# CHAPTER 1 INTRODUCTION

## 1.1 BACKGROUND OF STUDY

Customer satisfaction is a measure of how products or services provided by a company meet or exceed customer expectations. In a highly competitive market, any company can achieve a higher position by ensuring customer satisfaction (Leninkumar, 2017). Customer satisfaction can be the best predictor of whether or not they are likely to do business with a company again in the future. Thus, many studies have been done to uncover company’s service quality using customer feedback (Palese and Usai (2018); Sezgen et al. (2019); Chanwisitkul et al. (2018); Kim and Kang (2018); Wang et al. (2018)).

Customer reviews are information provided by the audience on how a company, product, or service assisted and met the needs of customers (Kierczak, 2020). The advent of technology has made it easier for businesses to obtain consumer feedback. Customers typically write about their experiences with a company’s product or service on review pages, community blogs, or forums (Ang, 2011). Feedback contains suggestions or complaints regarding a company, product, or service. Feedback identifies areas of a company’s product that can be improved, provides personalized recommendations, aids in the acquisition of new customers, or strengthens customer loyalty (Kierczak, 2020).

Utilizing feedback helps the company to make more profit or reduces costs. Understanding the root of the customers’ complaints can be a guide to develop market-oriented products or improve services (Joung et al., 2018). Furthermore, customer recommendations or expectations in reviews will assist the company in generating new ideas and improving product quality to optimize customer satisfaction. Therefore, managing and monitoring customer feedback is essential for businesses in today’s competitive marketplace.

Other than business, product or service reviews are also significant to other consumers. Besides commercials or advice from an expert, consumers seek an opinion from other buyers especially within the times of social media. People have more trust in customer reviews where other customers provide their viewpoint after purchasing or using a product and service. For instance, if a consumer is interested in a product, he or she will search reviews on a trustworthy blog or website to know whether the product is worth buying or not. Product review assists consumers to make wise choices based on ratings and comments from other reviewers.

Among all products, a smartphone is now essential to everyone. Smartphones and computer technology has spread rapidly around the globe with more people having smartphones and computers. A smartphone is a handheld device that combines the mobile telephone and computing functions into one unit. It typically has a touchscreen interface, internet access, and an operating system that is capable of downloading apps like a computer on top of being able to do the basic telephone capabilities such as messaging, making phone calls and taking pictures.

According to Statisa.com about the recent trends of customer purchases worldwide, the number of purchases of smartphones has increased significantly over the year of 2020 to 2021, with a staggering number of 260 million purchases, while the laptops have increased sales of 50 million, tablet sales are rather stagnant and personal desktop sales are decreasing. In addition, Google, on their website thingwithgoogle.com, published the fact that 66% of shoppers prefer to shop online. On the other hand, statcounter.com reported that the overall market share of mobile devices, desktop and laptop devices, and tablets, have the share of 55.44%, 41.91% and 2.65% respectively, showing that the mobile device industry has a bigger market share than all the other competing categories.

As the smartphone market is currently swarming with hundreds of products and keep on changing in technology, people are now concerned with issues related to smartphone users’ satisfaction. With the help of online reviews such as Amazon.com, the potential buyer may have a wider selection of options. Thus online review sites could offer much help in decision-making.

## 1.2 PROBLEM STATEMENT

Customers are increasingly opting for online shopping because it is convenient and costs savings (Miyatake, Nemoto, Nakaharai, & Hayashi, 2016). Besides that, online retailers allow consumers to leave feedback on items they’ve bought. Feedback from user reviews, especially the text that describes the experiences, comparisons, and features with a specific product may help other customers to make better decisions. It is also beneficial for businesses to identify their strengths and weaknesses and generate ideas to improve their product or services.

E-commerce platforms such as Amazon, eBay, Shopee or Lazada contain masses of product reviews that can be analyzed to gain valuable insight into customer needs. However, online customer reviews are in the form of unstructured text data that needs to be read to understand customers’ overall perception of a product or service. Customers were also asked to rate their overall experience, usually on a scale of 1 to 5, which may or may not reflect their feelings expressed in the textual feedback. Without a proper method, extracting meaningful information and useful knowledge from textual data is time-consuming and is a complicated process.

A method known as text mining has enabled researchers to analyze text data such as customer suggestions, customer complaints, and customer feedback (Joung et al., 2018). Thus with the aid of the text mining method, this study aims to uncover customer feedback from reviews in Amazon for selected smartphone products from March 2020 to June 2021.

## 1.3 RESEARCH OBJECTIVES

The research objectives of this study are:

1. To determine the frequent words used in online smartphone reviews.

2. To determine positive and negative reviews of smartphone models.

3. To determine the association between terms in reviews with features of smartphones.

4. To cluster online smartphone reviews.

## 

## 1.4 RESEARCH QUESTIONS

The research questions of this study are:

1. What are the frequent keywords used in online smartphone reviews?

2. What are the positive and negative reviews of smartphone models?

3. What are the strengths of the relationship between terms in reviews with features of smartphones?

4. How can the reviews of smartphones be categorised?

## 1.5 SCOPE AND LIMITATION OF STUDY

Like all other studies, this particular study has its limitations, both in terms of scope and methodology. Firstly, the study was carried out using reviews for items sold on Amazon.com and focused on specific smartphone brands and models through individual sales. Thus, the result might only be generalizable to the above population, meaning that findings might not apply to any other purchases that are made at other independent sellers on Amazon.com, other categories, or any other platforms.

Second, customer reviews are selected throughout March 2020 to June 2021, during the rise of the global pandemic of COVID-19, which forces consumers to buy their products off the internet, more than ever. Therefore, any data that is collected before or after the time mentioned might differ from our findings.

Thirdly, local online shopping sites such as Shopee.com.my and Lazada.com.my had an issue with the scraping process. Both sites have popup alerts when searching, hence, leave the choice to use a global online shopping site like Amazon.com. Thus, a newer result not only from Malaysia’s consumers but consumers from all corners of the globe.

Fourthly, since Amazon.com is a global online shopping site, it has not provided the option to filter out data by country, thus the reviews obtained were not limited to Malaysian. This is due to Amazon not requiring sellers to post country of origin, which circumvents statutes that require imported goods to display their point of origin, and offers no mechanism to filter such products despite years of complaints. Thus, this will result in data coming from all over the world.

Methodologically, this study will use R programming to determine the sentiment of the customer reviews. This study limits the language to be analysed only from English vocabulary, thus reviews containing non-English words will be filtered during pre-processing.

## 

## 

## 1.6 SIGNIFICANCE OF STUDY

The use of text mining to analyze customer reviews visualizes customer feedback more appealingly. Such findings help a company to identify suggestions or complaints regarding a company, product, or service much easier. Listening to customer feedback makes customers feel involved and important. Taking action on customer feedback will help the business improve its existing line of products and services. Besides that, customer feedback helps in decision making such as features in new product development or improving business marketing strategy. Lastly, this research will contribute to additional knowledge of related research fields.

# 

# CHAPTER 2 LITERATURE REVIEW

## 2.1 INTRODUCTION

In this chapter, previous studies by other researchers will be discussed. The first part explains the reason for customers and companies moving to e-commerce. Next, the second part elaborates on the analysis of online customer reviews.

## 2.2 MOVING TO E-COMMERCE

With the rapidly increasing number of people accessing the internet from day to day, added with the pandemic discouraging social interaction, it is very likely that online shopping is more preferable to traditional physical shopping. Four main factors that persuade people to shop online rather than physical shopping are ease, time-saving, convenience, and security (Chai & Yat, 2019). Other than that, shoppers’ lifestyles and technology development have also influenced people to shop online (Ibrahim et al., 2018). These factors are increasing as time passes, slowly but steadily replacing physical purchasing in favour of internet shopping.

Furthermore, due to the global pandemic, customers are reluctant to leave their homes and go to places where there is not enough social distance, so their preference has now shifted to online shopping. Consumers who had never shopped online found the internet as a marketplace and realised that it is the most secure choice to shop whatever they want, without getting out of their homes at this time (Grover, 2020). That is why e-commerce is gradually becoming more important as consumers continue to expect the same seamless, elevated experiences online as they would in person (Leahey, 2021).

This trend of decreasing in physical shopping and increasing in online shopping can already be seen in countries such as Italy and the UK, where the sales of physical stores have dropped by 7%, and 8% in Spain, while Amazon, has increased sales by 26% in the first quarter of 2020, compared to the same period in 2019. The pandemic has accelerated the adoption of digital by five years in only eight weeks, and this development does not appear to be slowing down anytime soon. The spike in sales is most apparent in the clothing, ornaments, accessories, house furnishings, and consumer electronics categories (Arora et al., 2020).

In the e-commerce industry, Amazon is one of the largest online retailers. Since its launch in 1995, the marketing strategy for Amazon has been to strengthen its brand, increase customer traffic to its websites, create customer loyalty, encourage repeat purchases and generate new product and service revenue (Majed et al., 2018). Now, Amazon.com is becoming a business model that adapts to different cultures (Garner, 2018).

Therefore, we wish to conduct analysis on customer reviews for certain products on Amazon. The products that will be analysed are smartphones with different models. The reason we chose to analyse on smartphones is because this product is important in our daily life. Smartphones allow us to call, text, send messages, shop, access the internet, and browse through our social media feeds (David et.al., 2021). Additionally, with the ongoing pandemic, the use of smartphones becomes even more critical. For example, Malaysians must use their smartphone to access the MySejahtera application to perform a scan when entering a shopping complex.

One of the most effective ways to build internet word-of-mouth is through online consumer review platforms ( C.R. Clark, U. Doraszelski, M. Draganska, 2007). As online word-of-mouth has grown in popularity, an increasing number of firms have begun to offer online word-of-mouth services. The vast consumer review systems on Amazon.com are well-known. The Internet has evolved into a very efficient and flexible platform for directly reaching people and creating buzz.

Product sales can be influenced by online user evaluations through either awareness or persuasion effects. Reviews express the presence of the product and so place it in the decision set of customers, according to awareness effects. Persuasive effects, on the other hand, are intended to alter customers' views and evaluations of a product, eventually influencing their purchasing choice. These two impacts have been extensively researched in previous marketing literature on the impact of advertising. Advertising has a strong beneficial influence on brand awareness, but no effect on perceived quality, according to the findings (C. Dellarocas, 2003).

## 2.3 THE ANALYSIS OF ONLINE CUSTOMER REVIEWS

Due to the advancement of technology, customers find that it is easier to post their experiences with products and services on the internet. Electronic Word of Mouth or eWOM refers to online customer review that provides information about many aspects of a product or a service, and that includes value for money and overall evaluation. The content of customer reviews can influence consumer buying behaviour. The overall rating of a product, as well as the average ratings of specific service or product features, are used by customers to obtain quantitative information on the quality and performance of the products or services. Online reviews remain important as they also focus on the qualitative dimensions of products or services (Chanwisitkul et al., 2018).

Customer reviews are an important source for businesses to analyze user requirements. Past studies showed that the use of text mining had assisted researchers to uncover important information from unstructured text from reviews. The method has been used in a study to analyze the online hotel review of ten hotels located in Bangkok. The study highlighted that friendly and helpful staff is the most influential factor affecting customer satisfaction. Meanwhile, the hotel’s complimentary facilities, such as the pool and Wi-Fi, have the greatest effect on customer dissatisfaction. Cleanliness, room and bathroom interiors, sleep quality, and location are among other factors that influence customer satisfaction. Text mining analysis helps to identify patterns in word frequencies while the regression analysis was carried out by using the overall customer satisfaction to verify the effects of service attributes, or factors (Chanwisitkul et al., 2018). A similar study was also done by Tian, He, Tao, and Akula (2016). Text mining and sentiment analysis on 58 hotels, rating from three to five-stars identified from TripAdvisor, had provided hotel managers with valuable management information and helped them to identify the strengths and weaknesses of their hotels.

Online reviews on e-commerce sites such as Amazon.com have a major impact on product reputation as they are heavily considered by potential buyers before making a purchase decision. Text mining techniques help to uncover customer attitudes and sentiments about products they have purchased and used. According to Jack and Tsai (2015), the method can compare and highlight top customer opinions of products and be used for future product improvement. Understanding the review sentiment which reflects the perceptions of a product, allows manufacturers to monitor the acceptance of their products. It also enables them to identify, fix, and resolve problems revealed in user reviews.

Meanwhile, Wang et al. (2018) had used text mining analysis and statistical modelling. They applied sentiment analysis to analyze online reviews of washing machines obtained from the Suning.com website to measure how the company’s attributes affect customer satisfaction. Then, a logistic regression model was developed to estimate the impact of various product properties on customer satisfaction scores. The results showed that drainage mode, loading type, frequency conversion, type, display, colour, and capacity affect customer satisfaction.

Nath et al. (2020) used sentiment analysis to analyse product online reviews and propose a method for clustering customer reviews for more informative summarization. Similar reviews were grouped based on the intensity of the customers' comments on key product attributes. The findings assist the manufacturer in identifying consumer satisfaction levels associated with various clusters and plan strategies to match those levels.

It is often that specific contexts are not considered and meaningful information is not extracted from the analysis of reviews. With text mining, through analysing the discriminative attributes of a competing product, more specific information can be derived than an overall product analysis. The analysis provides clarity on the strengths and weaknesses of competing products and provides realistic information that can help the company’s management activities (Kim & Kang, 2018).

Clustering is a prominent data mining approach that has been researched extensively in the context of text. It may be used in a variety of ways, including categorization, visualisation, and document organisation (Douglas Baker & Andrew Kachites McCallum, 1998). Clustering is the problem of locating groups of papers that are similar in a collection of documents. A similarity function is used to calculate the similarity. Text clustering may be done at various granularities, with clusters ranging from documents to paragraphs to phrases to words.

As a result, we'd like to attempt a few methods that have been used by other researchers. The descriptive, sentiment, correlation, and clustering analyses are used in our research on smartphone online reviews.

## 2.4 CONCLUSION

Online shopping is gradually replacing physical shopping as the new normal activity. It is caused by many reasons, but the existence of the COVID-19 pandemic greatly impacted the result. However, it remains unclear whether many people might return to traditional physical shopping or not when the pandemic is over.

Furthermore from the previous studies, online reviews are important because it helps a company to analyze customers’ feedback. In addition, it is also essential towards the development of qualitative products or services by looking at customer satisfaction through online reviews. Customer feedback on online reviews can be analyzed in many ways such as methods in text mining or statistical analysis.

Hence, we decided to conduct the methodologies used by many researchers on smartphone online reviews on Amazon.

# CHAPTER 3 METHODOLOGY

## 3.1 INTRODUCTION

This chapter discussed the research methods for the study. Research methodology can be defined as a set of systems, methods, procedures or rules that are used to conduct a structured research process for a thesis or dissertation. Thus, it allows for this study to obtain valid and reliable results that address the research aims and objectives. This chapter discussed the source of data and the text analysis method used in the study.

## 3.2 SOURCE OF DATA

The data source for this study is from an online shopping platform known worldwide, which is Amazon.com. The product reviews that were used for analysis were from the smartphone category. Reviews were collected from the top 4 products of the category, within the time range from March 2020 until June 2021. Through the data collection process, the data that is recorded is only the body of the review.

## 3.3 TEXT MINING ANALYSIS

Text mining, also known as text data mining, is a process of examining large collections of text and converting the unstructured text data into data for further analysis such as visualisation and model building (IBM, 2020). The structured data can also be used to identify meaningful patterns and insights. By applying advanced analytical techniques, researchers can explore and discover hidden relationships within the said data.

One of the advanced analytical techniques mentioned is Natural Language Processing (NLP), which helps machines to read and understand the text by simulating the human ability to understand a natural language such as English and Spanish (Milward, 2019). Text mining tools such as RStudio are capable of conducting the NLP analysis. The advantages of RStudio are it is free, reliable, and easy to use. The following chart in Figure 3.1 exhibits the process of text mining in this study. The process starts with data scraping from the website, importing text data to R, data cleaning, and analysis using descriptive statistics, sentiment analysis, finding the word association and performing cluster analysis.

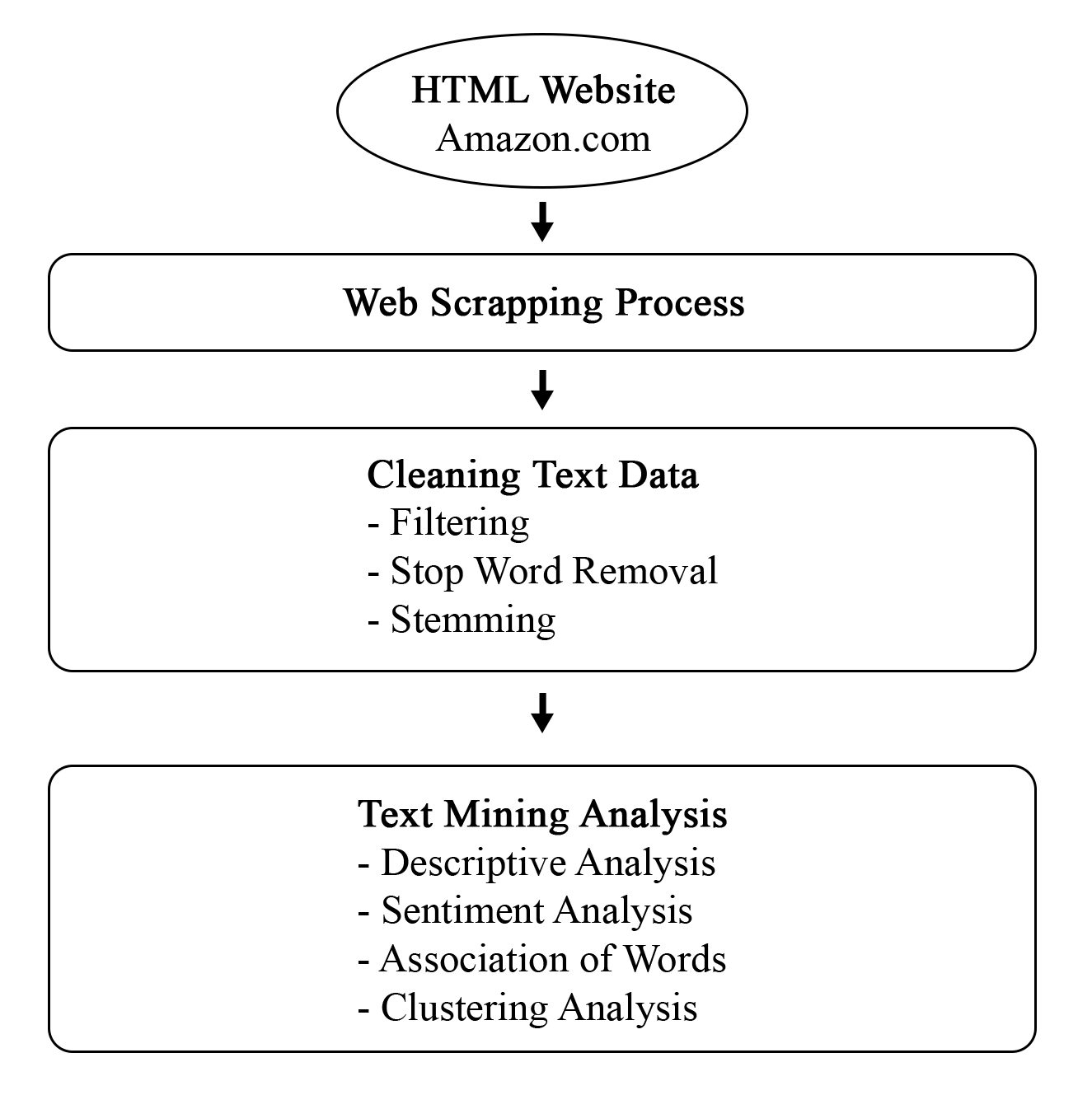


Figure 3.1: Text Mining Process

### 3.3.1 Web scraping process and tools

Web Scraping is the process of extracting content and data from a website. Web scraping using R mainly uses two libraries called *rvest* and *dplyr*. The *rvest* library helps to do the overall web scraping, while the *dplyr* library is just another tool that helps simplify the overall code.

The benefits of web scraping are it allows users to collect a huge amount of unstructured data autonomously and very fast (Saurkar et al., 2018). While the method of manually scraping the web is a viable option, when facing millions of data, having it done autonomously is a much more realistic method.

### 3.3.2 Cleaning text data

Cleaning text data is an extremely important process after data collection. This process can also be called text pre-processing. Data pre-processing is used to derive compelling and significant knowledge from unstructured text data in text mining (Kannan & Gurusamy, 2014). In addition, this process also removes unnecessary words, punctuation, numbers and many more. The common techniques in text data pre-processing are tokenization, filtering, stop word removal and stemming.

#### **3.3.2.**1 **Filtering**

According to Feinerer (2019), filtering is the process of removing documents from a database that meet certain criteria. The process allows users to use filter() to remove any terms with numeric digits, remaining punctuation, and missing observations.

#### **3.3.2.**2 **Stop word removal**

The purpose of this process is to remove common words from the file. Each word that is considered not vital should be eliminated. According to Welbers et al. (2017), filtering these terms out has the advantage of reducing data size, minimising computational effort, and even enhancing accuracy in some instances.

Stop words are common words that, in a natural language processing situation, do not provide much contextual meaning. These words are often the most common in a language. For example, the word “the”, “and”, “or”, etc. that have no meaningful significance in the study.

#### **3.3.2.**3 **Stemming**

Another method for text pre-processing is stemming, which is the procedure of detaching suffixes from words and deriving them to their root form (Lang, 2004). According to Sheela and Bharathi (2018), a stem is a group of words that are similar or nearly identical. In other words, it is a process that returns words into their common word. For instance, counting the words “reviewing”, “reviewer” and “reviewed” as being the same, which is “review”. The stemming process is essential and useful as it helps the programming to understand different words with the same meaning for a certain topic.

### 3.3.3 Term-document matrices

A term-document matrix is a mathematical matrix that depicts the frequency of terms found in a set of documents. Creating term-document matrices from a corpus is an ordinary method in text mining. Term-document matrix employs sparse matrices for corpora. The row in the matrix is filled with documents and each column represents a word or vice versa. It is used for the frequency of words occurring in the documents. This process is built into the *tm* packages in R programming.

### 3.3.4 Text mining analysis

#### **3.3.4.1** **Descriptive analysis**

Descriptive statistics summarize and organize characteristics of a data set either using numerical value or visualization. In text mining, ggplot and word cloud are some of the most common visualization methods used.

The ggplot from ggplot2 package is the most refined and satisfying design system accessible in R. The structure of the ggplot graph is pleasantly planned. It is a plotting package that simplifies generating complex plots from information in a data frame. Examples of plots that can be plotted using ggplots are scatter plots, box plots, and time series plots.

Meanwhile, generating word clouds involves two libraries in R which are wordcloud, and RColorBrewer. The purpose of word cloud is to provide a graphical presentation to determine the most frequent word in a document text. The size or colour of words in word clouds can be used to differentiate word frequencies. Generally, the greater word size indicates higher word frequency.

#### **3.3.4.2 Sentiment analysis**

Sentiment analysis is a method that identifies the hidden sentiment in a portion of words. It is a method of categorizing text to infer whether a section of text is positive or negative, or perhaps neutral. The *tidytext* package provides access to several sentiment lexicons. In this research, bing lexicon will be used because the bing lexicon categorizes words in a binary fashion into positive and negative categories.

Furthermore, the sentimentr package was used in R coding. The purpose of this package is to swiftly compute sentence polarity and to get the total sentiment measure for the whole review. Text polarity is the value of positivity or negativity, in each of the reviews.Then, the *ggplot2* package will be used to demonstrate the frequency of all the polarity values of reviews, to observe the overall trend of polarity in the reviews.

#### 3.3.4.3 Words Association

In addition to determining the frequencies of words and sentiments occurring within the reviews, there are also ways to understand the pairs of words that appear together in the reviews. According to Gulati, et.al. (2015), the approaches in this discipline are aimed at identifying patterns that determine the relationships between binary attributes (variables) used to characterise a group of objects.

Thus, this method allows this study to find the correlation between specific words as well as co-occurring of words in multiple documents. For example, if the words between *screen* and *scratch* have a high correlation, then both words have occurred together many times in a document. Not only can this process identify the co-occurring of words, but it can also give users more insights into the data. Therefore this study focused on words that highly correlated with three phone features which are battery, screen, and camera. R supports correlation analysis of text data in the *Ngram* package.

Statistically, Spearman’s Rank Correlation Coefficient was used to compare the association between words (Banerjee, 2003). This coefficient measures the correlation between two different rankings of a list of items. Specifically, given a set of Ngrams and their frequencies as observed in a corpus of text, it was ranked according to each of the two measures of association, and then compute the correlation between these two different rankings using following equation:

Where

n = total number of unique Ngrams in the corpus

Di = difference between the rankings of Ngram i in the two list

r = value of correlation

The value of r ranges from -1 to +1. A value of 0 implies no correlation between the two lists, while values that are further away from 0 imply greater correlation where the sign of the value indicates positive or negative correlation.

For this study, battery, screen, and camera are the chosen smartphone features for this study to see whether there were any fundamental words related to them. Words related with these features are believed to reveal some valuable insights. Hence, tm package of findAssocs() function was implemented in R programming. This function accepts three arguments: x, terms, and corlimit to measure word association norms using computational analysis (Nwofe & Goodall, 2017). In other words, the document term matrix, the term to discover correlations with, and the correlation limit represented as a number range from 0 to 1, are all delivered to the findAssocs() function. A value of 1 indicates that two words are constantly found together in documents, whereas a score around 0 indicates that the terms are rarely found together.

Since findAssocs() is calculated at the document level, every document that includes the term in question is associated with the other terms in that same document. However, documents that do not contain the search keyword are not considered. Moreover, a list of all other phrases that meet or exceed the minimum criteria will be returned by the findAssocs() function.

#### 3.3.4.4 Clustering Analysis

Clustering analysis is a statistical technique that is commonly used in solving classification problems. The purpose of this method is to gather various objects with similar attributes or traits together (Suzuki & Shimodaira, 2006). Hierarchical clustering is of two types. The first type is Agglomerative Hierarchical clustering, which starts at individual leaves and successfully merge clusters together. Commonly known as bottom up approach. Meanwhile Divisive Hierarchical clustering starts at root and recursively split the clusters, also known as top-down approach. Divisive approach is good for identifying large clusters while Agglomerative approach is good for identifying small clusters.

Some algorithm used for clustering are single linkage, complete linkage, average linkage (unweighted pair-group method), centroid method and Ward’s method. This study proposed agglomerative approach with average linkage algorithm where this algorithm will average the distances between all pairs.

Text clustering computes the similarity between text entities and forms clusters of entities. This study employed Euclidean distance to determine the similarity of text. Mathematically, the Euclidean distance is the straight line distance between two points in Euclidean space. It is measured by calculating the square root of the sum of squared differences between the two vectors elements. (Wang & Dong, 2020) defines the similarity measure using the using Euclidean distance



Hierarchical clustering generates a tree-like structure called Dendrogram. The Dendrogram branches out as it goes from the top to the bottom in the vertical direction, whereby the distance between the branches represents the distance between clusters. As the Dendrogram progresses down in its path, it splits the clusters down into smaller and smaller units until it reaches the data sample granularity level. And as vice versa, as it moves upwards, it will subsume smaller clusters into larger ones at each level until it reaches the entire system.

In R programming, the default combination of distance measure, dist() is Euclidan distance and algorithm hclust() is complete linkage. However, dist() proposes up to five other distance measures (maximum, manhattan, canberra, binary, and minkowski) and hclust() up to seven other algorithm methods (ward.D, ward.D2, single, average, mcquitty, median, and centroid). In this study, the list of words used for the clustering analysis was the top 20 words for each of the smartphone models. The R packages that were used to run the code are *stats* and *factorextra* package.

## 

## 

## 3.4 SUMMARY OF DATA ANALYSIS

The methodology for every objective is listed in Table 3.4, which summarises the data analysis.

Table 3.1 : Summary of Data Analysis

| NO | OBJECTIVES | VARIABLES | METHOD OF ANALYSIS |
| --- | --- | --- | --- |
| 1 | To determine the frequent words used across different smartphone models. | Body of reviews | Text mining analysis |
| 2 | To determine positive and negative reviews of smartphone models. |
| 3 | To determine the association between reviews. |
| 4 | To cluster the key attributes of smartphones. |

# CHAPTER 4 RESULT AND ANALYSIS

## 4.1 INTRODUCTION

This chapter reveals the results and analysis of the research derived from reviews in Amazon for four selected smartphone products for the year 2020 to 2021. The chapter is divided into four sections discussing findings for each objective. In this study, data was analysed using RStudio.

## 4.2 WORD FREQUENCIES OF TERMS IN SMARTPHONE REVIEWS

The following statistics in table 4.1 indicate summary statistics of terms frequencies throughout reviews. On the average, the frequency of a term used for this study is 14.6 times. Furthermore, the maximum frequency of a term is 1073 times, while some of the words are recorded as the lowest frequency, appearing only once in the smartphone online review. Additionally, the total number of words that emerge in the online review is 104,169.

Table 4.1: Statistics of terms in review

| **Statistics** | **Value** |
| --- | --- |
| Minimum frequency of a term | 1 |
| Maximum frequency of a term | 1,073 |
| Mean frequency of term | 14.6 |
| Total number of words | 104,169 |

### 

### 4.2.1 FREQUENCY OF TOP 20 TERMS IN ONLINE REVIEW

The top 20 terms that appear the most frequent in online review are explored in this section.

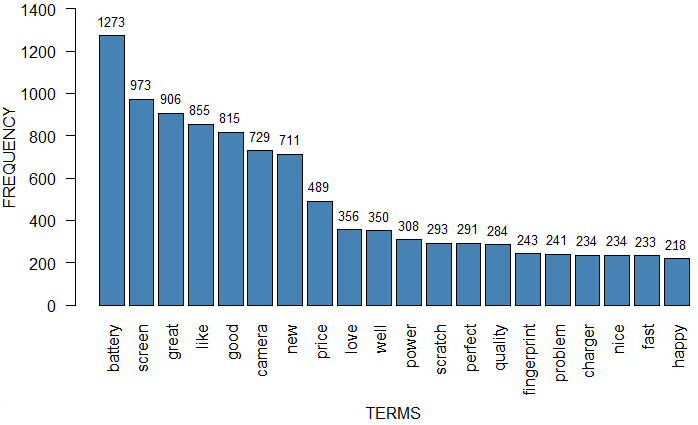


Figure 4.1: Bar plot of the top 20 terms in reviews

Based on the graph in figure 4.1, the word *battery* are the highest terms occurring in the customers’ smartphone reviews with frequency of 1,273. This result implies that most customers initially contemplate or inspect the battery when buying a smartphone. Thus, it shows that the battery features are essential to the customers in their reviews. Furthermore, the most common term after the *battery* is *screen* (973) and *great* (906), followed by *like* (855), *good* (815)*, camera* (729)and *new* (711).

### 

### 4.2.2 VISUALISING WORD FREQUENCY

The terms that appear the most frequent in online review are discussed in this section.

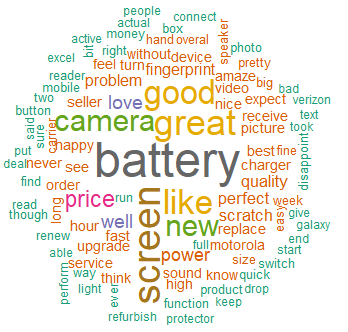


Figure 4.2: Word cloud of most frequent terms in the smartphone online reviews

Based on the word cloud displayed in figure 4.2, the word *battery* in the dark grey colour has the biggest size compared to all the other terms. Hence, *battery* is the most frequent word used in online reviews. Furthermore, the brown coloured term is the second largest in the word cloud. As a result, *screen* appears as the second-highest frequent term in the online review. Moreover, *good, great,* and *like* are the top five most frequent terms that emerged in the online reviewsince the yellow-coloured terms are the third largest in size. The next order of the most frequent terms is followed by the size of the coloured terms from largest to smallest; green, purple, red and blue.

This study reveals that the most frequent terms relate to the features of the smartphone such as battery, screen and camera, followed by positive terms such as great, good, like, happy, quality, upgrade and best, and some minor negative terms such as problem, disappointment, and bad. As mentioned by Lay-Yee et al. (2013) in their paper regarding “Factors Affecting Smartphone Purchase Decision among Malaysian Generation Y”, the most important aspect of purchasing a new device is regards to the features of a smartphone such as the battery, camera, and a bigger and brighter screen. Price plays a big factor in the selection process, to find the smartphone that suffices their need within their budget.

## 4.3 POSITIVE AND NEGATIVE TERMS IN REVIEWS ACROSS DIFFERENT MODELS

This section discussed the result of sentiment analysis using Bing Lexicon. Most common positive and negative terms across reviews of four smartphone models namely Google Pixel 4a, iPhone XR, Moto G Power and Samsung Galaxy S10 are presented in the following subsections.

### 4.3.1 Most Common Positive and Negative Terms in Google Pixel 4a Reviews

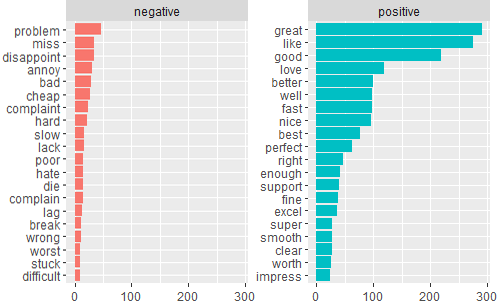


Figure 4.3: Horizontal Bar Chart of Positive and Negative Terms in Google Pixel 4a Reviews

Based on analysis, a total of 3181 words were associated with Google Pixel 4a’s feedback, with 2308 positive words and 873 negative words. According to figure 4.3, top 20 terms were displayed based on its category as positive and negative terms. The most used term for positive words is great followed by *like, good, love* and *better*. Meanwhile, the most common negative term that appears in reviews is *problem*. Other terms are *miss, disappoint* and *annoy*. Nevertheless, most customers stated that the smartphone model is a great product. Customer’s negative terms indicate dissatisfaction about the product’s performance and emotions such as disappointment, annoyance, and hate regarding the brand. Overall, customer reviews have been mostly positive based on the frequency of positive terms.

### 4.3.2 Most Common Positive and Negative Terms in iPhone XR Reviews

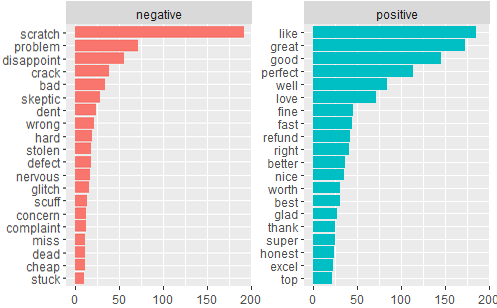


Figure 4.4: Horizontal Bar Chart of Positive and Negative Terms in iPhone XR Reviews

Reviews about smartphone model iPhone XR contains 2541 terms, with 1574 positive terms and 967 negative terms. The top 20 positive and negative terms are shown in figure 4.3. Term *like* is the most often positive term, while *scratch* is the most frequent negative term. In addition, customers show their satisfaction with the product in positive terms such as *great, good,* and *perfect*. On the other hand, terms such as *scratch, crack* and *dent* are some of the problems users face with the product. As a result, the response has been overwhelmingly positive.

### 4.3.3 Most Common Positive and Negative Terms in Moto G Power Reviews

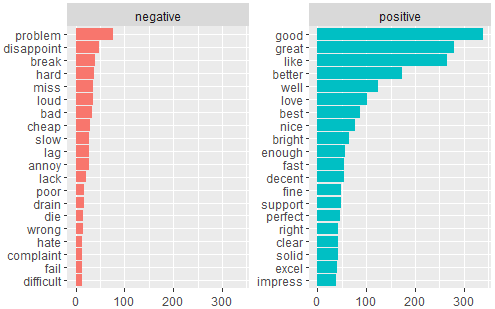


Figure 4.5: Horizontal Bar Chart of Positive and Negative Terms in Moto G Power Reviews

As illustrated in Figure 4.4, there are 3789 words relating to Moto G Power’s feedback, with 2710 positive words and 1079 negative words. The figure also includes the top 20 terms for each keyword. Term *good* is most commonly mentioned among positive terms, whereas *problem* is common in negative terms. Compared to other smartphone brands, there are less negative terms representing exterior failure such as scratch or dent. As a result, there has been a resoundingly favourable response.

### 4.3.4 Most Common Positive and Negative Terms in Samsung Galaxy S10 Reviews

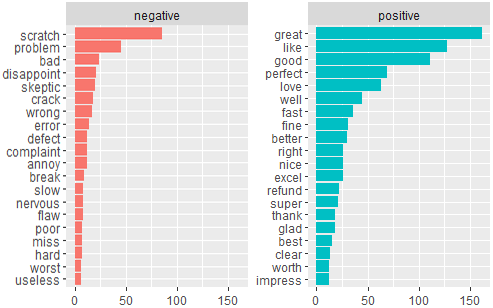


Figure 4.6: Horizontal Bar Chart of Positive and Negative Terms in Samsung Galaxy S10

There are a total of 1625 words related to the Samsung Galaxy S10’s feedback, as shown in Figure 4.5, with 1117 positive words and 535 negative words. The top 20 terms for positive and negative terms are shown in the graph. In the set of positive terms, *great* is the most common term, whereas *scratch* is utilised for negative terms. This shows that after using the brand, the majority of smartphone users believe it to be a great product. However, the scratch problem was always raised among users. As a result, there has been an overwhelmingly positive response for this brand.

## 

## 4.4 ASSOCIATION BETWEEN TERMS IN REVIEWS ACROSS DIFFERENT MODELS

Based on frequency analysis in 4.2, three smartphone features were identified as top terms in smartphone reviews. The three features mentioned are battery, screen and camera. Thus, this section intends to identify other terms associated with each smartphone feature.

### 4.4.1 Terms Associated with Battery

Table 4.2: Top 10 terms correlated with the term *battery*

| **No** | **Word** | **Correlation limit** |
| --- | --- | --- |
| 1 | brighter | 0.98 |
| 2 | demand | 0.98 |
| 3 | good | 0.98 |
| 4 | impress | 0.98 |
| 5 | pass | 0.98 |
| 6 | dual | 0.97 |
| 7 | end | 0.97 |
| 8 | focus | 0.97 |
| 9 | heavier | 0.97 |
| 10 | hold | 0.97 |

The term *battery* was found as the most frequent word in the reviews. Table 4.2 shows the top ten terms associated with *battery* with the correlation limit at least 0.97. The top five terms are *brighter, demand, good, impress,* and *pass*. As a proof, one of the examples that can be taken from the customer review that correlates with *brighter* is “if it were brighter, the battery likely wouldn't last as long”. In addition, the result shows that the terms that relate with batteries for all five smartphones are positive. In addition, the words associated with this feature are assumed to be pleasing and delightful for the manufacturer. Thus, motivating them in improving smartphone features in maintaining customers’ happiness.

### 4.4.2 Terms Associated with Screen

Table 4.3: Top 10 terms correlated with the term *screen*

| **No** | **Word** | **Correlation limit** |
| --- | --- | --- |
| 1 | eye | 1.00 |
| 2 | huge | 1.00 |
| 3 | app | 0.99 |
| 4 | average | 0.99 |
| 5 | camera | 0.99 |
| 6 | quit | 0.99 |
| 7 | bloatware | 0.98 |
| 8 | clear | 0.98 |
| 9 | feature | 0.98 |
| 10 | fit | 0.98 |

Based on table 4.3, the term *screen* has the highest correlation with the word *eye* and *huge*. That means *screen* occurred the most times together with the both words from the smartphone reviews. For instance in the online reviews, “my *eye*sight is aging and I need a bigger *screen*”. This result shows that most of the customers probably observe the size of a phone’s screen as it concerns the health of a human’s eyes. For example, a bigger screen on a smartphone may give people an advantage in getting more experiences in entertainment like playing games or watching videos on YouTube. On the other hand, plenty of today's portable gadgets for written communication like text messaging or email feature small screens, requiring close working distances and small font sizes (Rosenfield, 2011). These can put more constraints on optical accommodation and peripheral vision of a user, resulting in eye fatigue and headache (Mashalla, 2014).

### 4.4.3 Terms Associated with Camera

Table 4.4: Top 10 terms correlated with the term *camera*

| **No** | **Word** | **Correlation limit** |
| --- | --- | --- |
| 1 | command | 0.99 |
| 2 | exposure | 0.99 |
| 3 | fit | 0.99 |
| 4 | highest | 0.99 |
| 5 | launch | 0.99 |
| 6 | outstand | 0.99 |
| 7 | ratio | 0.99 |
| 8 | skin | 0.99 |
| 9 | sturdy | 0.99 |
| 10 | thinner | 0.99 |

Based on table 4.4, *camera* is highly correlated with many words that share the correlation limit of 0.99. For example, the top five terms associated with *camera* are *command, exposure, fit, highest* and *launch*. Furthermore, the result also shows that many customers are reviewing the functions of the smartphone’s camera. Some examples include command, which probably relates to the smartphone’s command to access its camera, and exposure or light that involves taking a photograph using a camera.

## 4.5 CLUSTERING OF TERMS IN REVIEWS ACROSS DIFFERENT MODELS

This section discusses the clustering of terms in online reviews across different models.

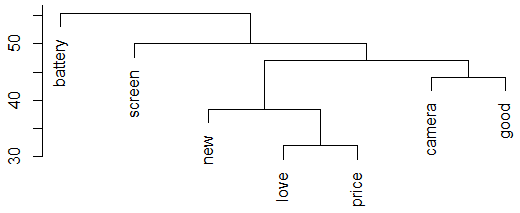


Figure 4.6: Dendrogram of terms in smartphone online review

Figure 4.6 shows the result of cluster analysis in the objective to identify the number of groups representing terms used in reviews. In addition, it may assist in classifying the topics discussed in the review.

The terms chosen in the analysis are the most frequent terms from all four smartphone model reviews. The dendrogram in figure 4.6 can be classified into four main clusters or groups. The first cluster consists of term *battery*, second cluster with term *screen*, the third cluster consists of *new, love* and *price*, and the fourth cluster consists of *camera* and *good*.

The first and second cluster is regarding the features of the smartphones, which are the screen and battery. The third cluster is grouped by customer emotion, love towards the smartphone brand, and price. The two terms *love* and *price* relate to *new* in its higher order. Finally, the final cluster is grouped by the feature deemed to be good, the camera.

These four clusters clearly show the topics discussed by most customers, the features and its price. This finding confirms that consumers will consider the features of a smartphone first, such as resolution of camera, the size and brightness of the screen and the battery life. The price of the smartphone also plays a major role in the selection of purchasing a new smartphone (Lay-Yee at al. ,2013)

# CHAPTER 5 CONCLUSION AND RECOMMENDATION

## 5.1 CONCLUSION

The main objective of this study is to determine the frequent words used across different smartphone models. There are two methods used to accomplish this objective, which are bar plot and word cloud. Not only do both methods obtain the same outcome, but they also complement each other’s result. The result shows that the phrase *battery* is the most frequently used in the smartphone online review with the frequency of 1,273. This indicates that most customers will review and share experiences on battery usage when purchasing a smartphone. Therefore, users in their reviews emphasise the importance of battery features.

Furthermore, the second objective is to determine positive and negative reviews for four smartphone models namely Google Pixel 4a, iPhone XR, Moto G Power and Samsung Galaxy S10. The result of this sentiment analysis is obtained by using Bing Lexicon. The most positive common term for Google Pixel 4a is *great*, while the most common negative term is *problem*. Meanwhile, *like* and *scratch* are both the most commonly positive and negative terms used for iPhone XR. As for Moto G Power, the term *good* appeared the most often in the online review while the term *problem* appeared the most for the negative. On the other hand for Samsung Galaxy S10, it is found that most of the customers used the term *great* for positive reviews, and the term *scratch* for negative reviews. Overall, all four smartphone models receive a higher number of positive responses from customers than the negative responses in the online review.

Moreover, the third objective for this study is to determine the association between terms in reviews with features of smartphones. The correlation method in R is utilised for this study to find the most associated terms with each three smartphone features: battery, screen, and *camera*. Terms that are the most highly associated with *battery* are *brighter, demand, good, impress* and *pass* with the correlation value of 0.98*.* As for *screen*, *eye* and *huge* are the top 2 terms with the highest correlation value of 1.00. Meanwhile for *camera*, all the top 10 terms are the most associated since they share the same correlation limit according to table 4.4.

The final objective for this study is to cluster online smartphone reviews. Dendrogram is used to achieve this objective. The terms included in the analysis are the ones that appear the most frequently across all four smartphone model reviews. The first cluster includes term *battery*, the second cluster includes term *screen*, the third cluster includes *new, love,* and *price*, and the fourth cluster includes *camera* and *good*. As a conclusion, all objectives are achieved.

Ultimately, all four smartphone models are good choices for purchasing due to the majority of positive customer ratings. This shows that these four smartphone models have met the majority of their customers' requirements. However, when evaluating customer sentiment in reviews, the Google Pixel 4a is the best smartphone model because it has the most positive terms and the fewest negative terms at the same time.

Aside from that, when developing a smartphone, features such as battery, screen, camera, and others must be prioritised. It is the key factor in whether or not a smartphone can meet the needs of its users. For instance, customers may look for features such as a long-lasting battery, a large screen with no scratches, or a high-quality camera.

## 5.2 RECOMMENDATION

This research aims to retrieve information from online reviews for the smartphone category in Amazon.com. The study's objectives were obtained primarily to the utilisation of a specific methodology. However, there are still drawbacks in terms of data sources, and analysing several sources at the same time, as well as the visualisation of the analysis. Several suggestions have been made to help upgrade the effectiveness of future research.

To acquire insights from similar data, it is recommended that people employ alternative packages of R programmes for further analysis. In terms of data exploration, R software provides vast packages of statistics and data mining. A longer time range for data collecting, on the other hand, should be considered for future research, allowing for a larger sample size to be evaluated. Finally, because there are many companies that are not identical to Amazon.com, researchers may acquire data from a number of platforms. Additionally when carrying out a research regarding Malaysia, researchers must consider platforms from Malaysia only.

When conducting text mining analysis, researchers can use the Pearson and Spearman methods, as well as Kendall correlation, to examine associations. The corr.test () function in the psych package is the best way to use these methods. Furthermore, one of the alternatives for extracting collocation from data is to create an n-gram list in the bag-of-words by using the unnest tokens() function in the tidytext package. In comparison to the corlimit function, the outcome of this function can provide researchers with more insights and a broader point of view.

Furthermore, users should carefully consider before purchasing a smartphone. Customers should pay closer attention to smartphone features like battery, screen, and camera, according to this survey. As these features were highlighted the most in reviews compared to other features, many customers perceive that these are the deciding factor in whether or not a smartphone is a good or bad product. Besides that, customers also must focus on the online reviews in order to learn more about the details and gain valuable insights on which smartphone can provide them with the most satisfaction.

## 

## REFERENCES

Ang, L. (2011). Community relationship management and social media. *Journal of Database Marketing & Customer Strategy Management*, *18*(1), 31–38.

Arora, A., Dahlstro¨m, P., Hazan, E., Khan, H., & Khanna, R. (2020). Reimagining marketing in the next normal. *McKinsey & Company, Marketing & Sales*.

Banerjee, S., & Pedersen, T. (2003, February). The design, implementation, and use of the ngram statistics package. In *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 370-381). Springer, Berlin, Heidelberg.

Chai, L. T., & Yat, D. N. C. (2019). Online food delivery services: making food delivery the new normal. *Journal of Marketing advances and Practices*, *1*(1), 62–77.

Chanwisitkul, P., Shahgholian, A., & Mehandjiev, N. (2018). The reason behind the rating: Text mining of online hotel reviews. In *2018 ieee 20th conference on business informatics (cbi)* (Vol. 1, pp. 149–157).

David, M. E., & Roberts, J. A. (2021). Smartphone use during the COVID-19 pandemic: Social versus physical distancing. *International Journal of Environmental Research and Public Health*, *18*(3), 1034.

Feinerer, I. (2013). Introduction to the tm Package Text Mining in R.

Garner, B. A. (2018). Amazon in the global market. *Journal of Marketing and Management*, *9*(2), 63–73.

Grover, S. (2020). Covid-19 aftermath: Online retail shopping becoming the new normal. *EntreprenuerIndia*.

Gulati, H., & Singh, P. K. (2015, March). Clustering techniques in data mining: A comparison. *In* *2015 2nd international conference on computing for sustainable global development (INDIACom)* (pp. 410-415). IEEE.

Hsin-KaiYu & I-ChinWu (2020). A recommender system using GA K-means clustering in an online shopping market.

IBM (2020). What is Text Mining?. *IBM Cloud Learn Hub.*

Ibrahim, I. I., Razali, M. F. M., Zulkifli, S. A., Nor, M. N. M., Mansor, Z., Nasir, N. F. M., & Kassim, K. M. (2018). Determinants that persuading students’online shopping behavior towards purchasing cosmetic product. *International Journal*, *1*(5), 67–75.

Jack, L., & Tsai, Y. (2015). Using text mining of amazon reviews to explore user-defined product highlights and issues. In *Proceedings of the international conference on data science (icdata)* (p. 92).

Joung, J., Jung, K., Ko, S., & Kim, K. (2018). Customer complaints analysis using text mining and outcome-driven innovation method for market-oriented product development. *Sustainability*, *11*(1), 1–14.

Kannan, S., & Gurusamy, V. (2014). Preprocessing techniques for text mining. *International Journal of Computer Science & Communication Networks*, *5*(1), 7-16.

Kierczak, L. (2020). Customer satisfaction: Why it’s still important in 2021. *Journal of Air Transport Management*.

Kim, S. G., & Kang, J. (2018). Analyzing the discriminative attributes of products using text mining focused on cosmetic reviews. *Information Processing & Management*, *54*(6), 938–957.

Lang, D. T. (2004). *Word* *Stemming in R. Department of Statistics, UC Davis.*

Leahey, L. (2021). The “new normal” of online shopping. *Cyncity*.

Leninkumar, V. (2017). The relationship between customer satisfaction and customer trust on customer loyalty. *International Journal of Academic Research in Business and Social Sciences*, *7*(4), 450–465.

Majed, S. Z., Nuraddin, S. H., & Hama, S. V. S. (2018). Analyzing the amazon success strategies. *Journal of process management. New Technologies*, *6*(4), 65–69.

Mashalla, Y. J. (2014). Impact of computer technology on health: Computer Vision Syndrome (CVS). *Medical Practice and Reviews*, *5*(3), 20-30.

Milward, D. (2019) What is Text Mining, Text Analytics and Natural Language Processing? *Linguamatics.*

Miyatake, K., Nemoto, T., Nakaharai, S., & Hayashi, K. (2016). Reduction in consumers’ purchasing cost by online shopping. *Transportation Research Procedia*, *12*, 656– 666.

Mullen, L. A., Benoit, K., Keyes, O., Selivanov, D., & Arnold, J. (2018). Fast, consistent tokenization of natural language text. *Journal of Open Source Software*, *3*(23), 655.

Nath, G., Ghosh, R., Risk, R., & Nath, R. *Cluster Analysis of Customer Reviews: Summarizing Customer Reviews to Help Manufacturers Identify Customer Satisfaction Level.*

Palese, B., & Usai, A. (2018). The relative importance of service quality dimensions in e-commerce experiences. *International Journal of Information Management*, *40*, 132–140.

Rosenfield, M. (2011). Computer vision syndrome: a review of ocular causes and potential treatments. *Ophthalmic and Physiological Optics*, *31*(5), 502-515.

Saurkar, A. V., Pathare, K. G., & Gode, S. A. (2018). An overview on web scraping techniques and tools. *International Journal on Future Revolution in Computer Science & Communication Engineering*, *4*(4), 363-367.

Sezgen, E., Mason, K. J., & Mayer, R. (2019). Voice of airline passenger: A text mining approach to understand customer satisfaction. *Journal of Air Transport Management*, *77*, 65–74.

Sheela, S., & Bharathi, T. (2018). *Analyzing different approaches of text mining techniques* *and* applications*. International Journal of Computer Science Trends and Technology, 6(4), 23–29.*

Suzuki, R., & Shimodaira, H. (2006). Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics, 22(12), 1540-1542.

Tian, X., He, W., Tao, R., & Akula, V. (2016). Mining online hotel reviews: a case study from hotels in china.

Wang, Y., Lu, X., & Tan, Y. (2018). Impact of product attributes on customer satisfaction: An analysis of online reviews for washing machines. *Electronic Commerce Research and Applications*, *29*, 1–11.

Welbers, K., Van Atteveldt, W., & Benoit, K. (2017). *Text analysis in r. Communication Methods and Measures, 11(4), 245–265.*

C.R. Clark, U. Doraszelski, M. Draganska, Information or persuasion? An empirical investigation of the effect of advertising on brand awareness and perceived quality using panel data, Stanford University Graduate School of Business Research Paper, 2007, No. 1971

C. Dellarocas, The digitization of word of mouth: promise and challenges of online feedback mechanisms, Management Science 49 (10) (2003) 1407–1424

**APPENDIX A**

**Research Schedule**

Research schedule

|  | 2021 | | | | | | | | | 2022 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Activity | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov | Dec | Jan |
| Planning research report |  |  |  |  |  |  |  |  |  |  |
| Brainstorming topic |  |  |  |  |  |  |  |  |  |  |
| Choose research topic |  |  |  |  |  |  |  |  |  |  |
| Problem statement, research objective, research question |  |  |  |  |  |  |  |  |  |  |
| Introduction |  |  |  |  |  |  |  |  |  |  |
| Completing scope and limitation |  |  |  |  |  |  |  |  |  |  |
| Literature review |  |  |  |  |  |  |  |  |  |  |
| Methodology |  |  |  |  |  |  |  |  |  |  |
| Collection of data |  |  |  |  |  |  |  |  |  |  |
| Identifying veriables |  |  |  |  |  |  |  |  |  |  |
| Analysing data |  |  |  |  |  |  |  |  |  |  |
| Data interpretation |  |  |  |  |  |  |  |  |  |  |
| Writing up research report |  |  |  |  |  |  |  |  |  |  |
| Presentation and submit full research report |  |  |  |  |  |  |  |  |  |  |

**APPENDIX B**

library(tm)

library(tidytext)

library(widyr)

library(stringr)

library(tidytext)

library(tidyverse)

library(wordcloud)

library(SnowballC)

library(RColorBrewer)

docs<- Corpus(DirSource('/Users/admin/Desktop/fyp 100'))

inspect(docs)

**FILTERING**

docs2 <- tm\_map(docs2, removePunctuation)

docs2 <- tm\_map(docs2, content\_transformer(tolower))

docs2 <- tm\_map(docs2, removeNumbers)

docs2 <- tm\_map(docs2, stripWhitespace)

**STOPWORDS**

docs2 <- tm\_map(docs2, removeWords, stopwords("english"))

**STEMMING**

*docs2 <- tm\_map(docs2, stemDocument)*

**TERM-DOCUMENT MATRIX**

*dtm <- DocumentTermMatrix(docs4)*

*freq <- colSums(as.matrix(dtm)) #collapse matrix by summing over columns*

*length(freq) #length should be total number of terms*

*dtmr <- DocumentTermMatrix(docs4, control = list(wordLength = c(4,20), bounds = list(global = c(3,27))))*

*freqr <- colSums(as.matrix(dtmr))*

*length(freqr)*

*ordr <- order(freqr, decreasing = TRUE) #create sort order (asc)*

*freqr[head(ordr)] #inspect most frequently occurring term*

*freqr[tail(ordr)] #inspect least frequently occurring terms*

*#list the most frequent terms. Lower bound specified as second argument*

*findFreqTerms(dtmr, lowfreq = 80)*

**BAR PLOT**

**WORDCLOUD**

wordcloud(corpus4)

wordcloud(corpus4, min.freq = 1, random.order = FALSE, col = rainbow(5)) #combininng all documents

comparison.cloud(corpus4)

corpus5 <- Corpus(VectorSource(corpus4))

tdm <- TermDocumentMatrix(corpus5) #to see the highest frequency in each document

m <- as.matrix(tdm) #the list of words frequency for each document

colnames(m) <- c("HEADPHONE", "SOFA") #change column name

comparison.cloud(m)

**SENTIMENT ANALYSIS**Example for model Google Pixel 4a (gp4)  
Reference: https://www.tidytextmining.com/tidytext.html

1. For creating top 20 positive and negative words  
  
gp4 <- readLines("Google\_Pixel\_4a.txt")

gp4\_df <- tibble(line = 1:506, gp4 = gp4) *#create tidytext(df=dataframe), 506=no. of columns*gp4\_df <- gp4\_df %>%  
 unnest\_tokens(word, gp4) %>%  
 ungroup()

sentimentipx <- ipxr\_df %>% *#scanning for sentiment words*  
 inner\_join(get\_sentiments("bing")) %>%  
 count(word, sentiment, sort=TRUE) %>%  
 ungroup()

sentimentipx %>% *#plotting*  
 group\_by(sentiment) %>%  
 slice\_max(n, n = 20) %>%   
 ungroup() %>%  
 mutate(word = reorder(word, n)) %>%  
 ggplot(aes(n, word, fill = sentiment)) +  
 geom\_col(show.legend = FALSE) +  
 facet\_wrap(~sentiment, scales = "free\_y") +  
 labs(x = "Contribution to sentiment",  
 y = NULL)

2. To find sum of positive & negative words  
Reference: https://www.youtube.com/watch?v=WfoVINuxIJA

poswords<- scan('poswords.txt', what='character', comment.char=';') *#bag of positive words*

negwords<- scan('negwords.txt', what='character', comment.char=';') *#bag of negative words*

bwgp4<-str\_split(gp4, pattern = "\\s+") *#bw = bag of words, creating a bag of words*

bwgp4 <- unlist(bwgp4) *#convert from list to non-list type document*

sum(!is.na(match(bwgp4, poswords))) *#find sum of positive words*

sum(!is.na(match(bwgp4, poswords))) *#find sum of negative words*

**ASSOCIATION OF WORDS**

*wordasso <- findAssocs(dtmr,"battery",corlimit = 0.9)*

*wordasso <- findAssocs(dtmr,"screen",corlimit = 0.9)*

*wordasso <- findAssocs(dtmr,"camera",corlimit = 0.9)*

**CLUSTERING**

*res.hc <- hclust(d = res.dist, method = “moderate”)*

*plot(x = res.hc)*

*fviz\_dend(x = res.hc, cex = 0.7, lwd = 0.7)*

jj <- str\_split(corpus4, pattern = "\\s+")

lapply(jj, function(x){sum(!is.na(match(x, opinion.lexicon.pos)))})

lapply(jj, function(x){sum(!is.na(match(x, opinion.lexicon.neg)))})

lapply(jj, function(x){sum(!is.na(match(x, opinion.lexicon.pos))) - sum(!is.na(match(x, opinion.lexicon.neg)))})

score <- unlist(lapply(jj, function(x){sum(!is.na(match(x, opinion.lexicon.pos))) - sum(!is.na(match(x, opinion.lexicon.neg)))}))

mean(score)

sd(score)

hist(score)

**WEB SCRAPING**

reviewlist = data.frame()

for (pg\_no in seq(from = 1, to = 3, by = 1)){

link = paste0("Insert Amazon link here")

page = read\_html(link)

review = page %>% html\_nodes(".review-text-content span") %>% html\_text()

reviewlist = rbind(reviewlist, data.frame(review))

PEMBETULAN

~~1. Review and feedback, sama ke tidak? Kalua sama, seragamkan semua review sahaja~~

2. LR kurang discussion on findings, especially clustering.

~~3. "user" dalam chap 4 macam overall pembeli smartphone. bukan pembeli of the 4 models.~~

4. chap 5 bukan ringkasan chap 4 sebenarnya,. next time try to compare all 4 findings. what are the best information that can be revelaed.

~~5. what is the recommendation that i can get as a buyer? apa boleh dapat dari all this analysis. Not mentioned,, no conclusive statement~~

6. kenapa frequency 4.2 across all models, tapi; 4.3 specific by model?

~~7. abstract tak mention the 4 clusters yang dapat.~~

8. LR 2.2 "moving to ecommerce" tak penting. patutnya cakap pasal "online review" lagi. what is the advantage, platform, how to make use

9. LR, eg, 2.3, at the end of the section kena ada conclusion of what they do. "many researcher use sentiment analysis. thus this inspired us to use it as well". jangan just list down. dari LR tu kena ada relate dengan kita punya paper as well, kita punya comment as well

~~10. 3.2 source of data, patutnya kena mention how many reviews yang kita amik. to ensure that the comparison semua adil~~

11. method of analysis. method of analysis patutnya ada mention reason. descriptive. why descriptive? it should refer back to objective

~~12. table in 3.4 patutnya kena ada explaination. tak boleh ada table, and thats it~~

~~13. 4.2 maximum term (1073) tak sama dengan 4.2.1. (1273),, and then maximum tu apa? too technical, should put “most frequent word” or “word with the highest frequency”~~

~~14. 4.2.1 bila ada section, patutnya ada text dulu. baru graph, secara ringkas. Tak boleh section, then terus graph~~

~~15. bila rujuk "table.3.3" , "figure 2.1", T dengan F kena besar~~

16. patutnya kasi contoh of reviews dalam 4.2.2. kalau great, pasal great apa ? same case dalam 4.4. how is "battery" related to "bright"

17. dendrogram takleh pakai sebab pelik. its not conclusive

18. for association, tulis spearman, tapi assoch bukan chi-square? source of data tu explain about structure. lepas dah clean apa semua tu, how dia jadi ranking? sebab text ni categorical. so bila categorical dia jadi chi square/pearson.

Pembetulan yang dah buat yang panel tunjuk dalam returned document

**Dr Nisa;**

* Typo errors

**Pn AIni**;

* alkdjfla