

**UNIVERSITY OF
WESTMINSTER**

**SCHOOL OF
COMPUTER SCIENCE & ENGINEERING**

**STAR RATING PREDICTION USING SENTIMENT
ANALYSIS OF AMAZON MOBILE PHONE DATASET**

**BY
RAJ PRAVIN RAJENDRAN
(W1795435)**

Supervised by
ROLF BANZIGER


Submitted in partial fulfilment of the requirements of
the School of Computer Science & Engineering
of the University of Westminster
for award of the Master of Science

SEPTEMBER 2022

DECLARATION

I, **RAJ PRAVIN RAJENDRAN**, declare that I am the sole author of this Project; that all references cited have been consulted; that I have conducted all work of which this is a record, and that the finished work lies within the prescribed word limits.


This has not previously been accepted as part of any other degree submission.

Signed : 

Date : 08/09/2022

FORM OF CONSENT

I, **RAJ PRAVIN RAJENDRAN**, hereby consent that this Project, submitted in partial fulfilment of the requirements for the award of the MSc degree, if successful, may be made available in paper or electronic format for inter-library loan or photocopying (subject to the law of copyright), and that the title and abstract may be made available to outside organisations.

Signed : 

Date : 08/09/2022

TABLE OF CONTENTS

1. ABSTRACT	6
2. INTRODUCTION.....	6
3. LITERATURE REVIEW	7
3.1 GLOBAL MOBILE PHONE MARKET.....	7
3.2 THE EVOLUTION OF THE MOBILE PHONE INDUSTRY	8
3.3 KEY BRAND IN MOBILE PHONE INDUSTRY	9
3.4 WHY WOULD AMAZON WANT TO SELL A MOBILE PHONE.....	9
3.5 TOPIC MODELLING	11
3.6 SENTIMENT ANALYSIS	12
3.7 WHY SENTIMENT ANALYSIS IS IMPORTANT	12
3.8 SENTIMENT ANALYSIS APPROACHES	12
3.9 UNSUPERVISED LEARNING APPROACH	13
3.10 SUPERVISED LEARNING APPROACH	13
3.11 WORD 2 VECTOR	13
4. PROBLEM STATEMENT	14
5. AIM AND OBJECTIVES.....	14
5.1 AIM.....	14
5.2 PRIMARY OBJECTIVES	14
5.3 SECONDARY OBJECTIVES	15
6. METHODOLOGY	15
7. TOOL CHOSEN.....	15
7.1 PYTHON	15
7.2 JUSTIFICATION TO USE PYTHON.....	16
8. RESEARCH PROJECT DESIGN STAGES	17
8.1 PROJECT FLOWCHART	17
8.2 STEPS OF THE PROJECT	18
9. PROJECT PROGRAMMING LANGUAGE AND ENVIRONMENT	18
9.1 DATA COLLECTION	19
9.1.1 REVIEWS DATASET.....	19
9.1.2 PRODUCT DATASET	20
9.2 EXPLORATORY DATA ANALYSIS	20
9.3 DATA PRE-PROCESSING	20
9.4 DATA VISUALISATION	21
9.5 MODEL CREATION	21
9.6 MODEL AUTOMATION	21
9.6 HOW MY PROJECT IS DIFFERENT FROM OTHERS	21

10. RESEARCH PROJECT STAGES IMPLEMENTATION	22
10.1 IMPORTING THE DATA IN JUPYTER NOTEBOOK	22
10.2 DATA TYPES	22
10.3 DROPPING THE UNIMPORTANT COLUMNS	23
10.4 IDENTIFICATION AND TREATMENT OF MISSING VALUES	24
10.5 DATA MERGING	24
10.6 STATISTICAL SUMMARY	25
10.7 EXPLORATORY DATA ANALYSIS	25
10.8 ASSIGNING THE TARGET VARIABLE	27
10.9 WHY TARGET VARIABLE IS IMPORTANT	27
10.10 TARGET VARIABLE ANALYSIS	28
10.11 LOWERCASING THE STRING DATA	29
10.12 REMOVING THE PUNCTUATION	30
10.13 REMOVING STOP WORDS	30
10.14 TRANSFORMING "BODY" USING TF-IDF VECTOR	31
10.15 TRAIN TEST SPLIT	33
10.15.1 WHEN TO USE TRAIN-TEST SPLIT	33
10.15.2 HOW TO CONFIGURE THE TRAIN-TEST SPLIT	34
10.16 MODEL CREATION	35
10.16.1 LOGISTIC REGRESSION	35
10.16.2 NAÏVE BAYE'S CLASSIFIER	38
10.16.3 SUPPORT VECTOR MACHINE	39
10.16.4 DECISION TREE CLASSIFIER	40
10.16.5 RANDOM FOREST CLASSIFIER	42
10.16.6 ADA BOOSTING	44
10.17 AUTOMATING THE MODEL	45
11. RESULT AND OUTCOME	47
12. CHOOSING THE MOST EFFECTIVE MODEL	47
13. CONCLUSION	48
13.1 RESEARCH PROJECT OBJECTIVES	48
13.2 PROJECT PROBLEM STATEMENT	48
13.3 EVALUATION	48
14. FUTURE WORK	48
15. REFERENCES	49

1. ABSTRACT:

Amazon links customers with fantastic local businesses. In this article, we will concentrate on mobile phone brand reviews. We want to estimate a mobile phone's rating based on past information such as the title, and the Amazon user's feedback. We explore the data set made available by Kaggle. We will forecast the star(rating) of a review in this project. Logistic regression, Random Forest, SVM, Decision tree, ADA boosting, and naive bayes model are used in conjunction with sentiment analysis. After analyzing the performance of each model, the SVM algorithm was determined to be the best model for predicting ratings from reviews. In addition, we automated the model creation part In python for the computer efficiency which helps in saving time.

2. INTRODUCTION:

Since 2008, the Mobile phone market has been gradually expanding, both in terms of market size and the number of devices and vendors. Global shipments of mobile phones are anticipated to reach 1.43 billion units in 2022.

With a stunning 39.1% growth from its market share of 52.4 percent in 2016 to its current market share of 72.9 percent, mobile ecommerce sales now account for a larger portion of overall ecommerce sales. To put it another way, today, mobile devices account for about three out of every four dollars spent on online sales.

People make online mobile phone purchases and offer feedback to the specific e-commerce sector. The e-commerce sector can examine client comments and present a solution that surpasses the complaint. Additionally, this project's predictions will enable businesses to make extremely accurate assumptions about the most likely outcomes of a question based on historical data. These assumptions can be made about a variety of topics, including the likelihood of customer churn, potential fraudulent activity, and more.

Since **Amazon** is the largest industry, I chose it as my e-commerce sector since the data there would include a large number of mobile phone brand names, and the data instances would be enormous. Furthermore, if we have a large amount of data, we can easily improve the model's accuracy, which is crucial for almost all machine learning techniques.

A global American technology business, **Amazon.com**, Inc., specializes in e-commerce, cloud computing, digital streaming, and artificial intelligence. One of the most valuable brands in the world, it has been called "one of the most significant economic and cultural forces in the globe along with Alphabet, Apple, Microsoft, and Meta, it is one of the Big Five American technological firms.

The goal of this project is to enable Amazon.com to obtain a knowledge of their customers' attitude toward mobile phone products, service, and the organization as a whole, as well as to predict

sentiment for data set. This will be accomplished by developing and deploying a sentiment analysis tool that is linked to the present CRM system and uses feedback and star rating data as input. I feel that by collecting this data, the company will be able to make better product selections and respond to negative comments more quickly, resulting in greater client relationships and retention.

Additionally, my deliverable is to find the sentiment score for certain brand and finally predict the star ratings whether the brand gets right reviews on the catalogue page or not and to predict what rating would the user give for a specific brand

This study presented a mechanism that might provide a quantitative perspective of the individuals purchasing the mobile phone brands and offering feedback about it in order to analyze all of this unstructured Amazon data. The technology recognized reputational strengths and weaknesses and offered insights into customer reviews of Amazon. This study aims to add to the body of knowledge on e-reputation measurement since there is a dearth of research in the academic community.

Furthermore, creating a Machine learning pipeline in terms of automating the entire ML features using the library **Tkinter**.

The Python package Tkinter is frequently used to build graphical user interface (GUI) applications. The process of creating a GUI with Tkinter is quite simple and quick. Several widgets from Tkinter can be utilized for creating a graphical user interface. These consist of buttons, checkboxes, radio buttons, etc. After developing the machine learning model later in the article, we will examine how to create a GUI using Tkinter. Finally, we can determine which model performs the highest accuracy with the use of the Tkinter GUI, and with the aid of the GUI, we can freeze the model and deploy it on cloud platforms.

3. LITERATURE REVIEW

3.1 GLOBAL MOBILE PHONE MARKET

In 2014, the number of mobile subscribers increased globally by 5%. As penetration rates get closer to saturation levels, developed market growth slows. For instance, unique subscriber growth in Europe and North America in 2014 was around 1%. On the other end of the spectrum, Sub-Saharan Africa continued to be the region with the lowest penetration rates in the world, with subscriber growth at just under 12%.

Due to factors like decreased costs, improved designs and functionalities like improved mobile browsing and email services, the emergence of new network technologies like 3G and 4G, improved professional and personal data supervision, and the standardization and upgrading of all operating systems, the market for mobile phones and smart phones is currently experiencing proliferation.

It is tough for vendors to maintain their market shares in this highly competitive industry where major firms face fierce competition from regional players. For instance, over the last two years, Nokia has lost a sizable portion of its market share. Some of the operating systems used in smartphones include Android, iPhone OS, BlackBerry OS, Symbian, and Windows. In North America, the Blackberry operating system is well-liked. In North America, the iPhone operating system has recently experienced rapid expansion, and this growth is expected to continue in the coming years. The market is expanding as a result of increased internet usage brought on by technical improvements and network infrastructure upgrades.

3.2 THE EVOLUTION OF THE MOBILE PHONE INDUSTRY:

The mobile phone market is very dynamic. The industry won't disintegrate, but it's very obvious that things will be very different in the future. In fact, you probably wouldn't recognize the market if you were to take a look at it in ten years.

- Marty Cooper, a Motorola engineer, placed the first call using a "true handheld portable cell phone" in April 1973.
- A radio common carrier (RCC), a system that predated cellular technology and was established in the 1960s. It had its own phone number and could send voice communication using a push-to-talk technology, similar to a radio, but it made use of the public telephone network.
- The clamshell form factor was the first step toward completely portable devices.
- Nokia was a pioneer in this field of technology. Because it resembled a candy bar in terms of size and shape, the candy bar phone was given that name.
- The mobile industry underwent change in the middle of the 1990s. The current flip phone was made possible by the clamshell phone, which lightened its load.
- An international call can be placed using a satellite phone, which connects to orbiting satellites rather than cellular towers on the ground.
- The personal digital assistants (PDA) of the 1990s heralded the arrival of a flurry of touchscreen and pocket computing gadgets. With the introduction of the Palm Pilot in 1997, Palm popularized the industry game changer.
- The widely used Nokia 6000 Series made mobile communication accessible and broadly cheap for the general public in the early 2000s.

- The Motorola Flip Phone, introduced in 2004, has a form factor known as the Razr that is extremely small, sleek, and portable.
- When it first debuted in the early 2000s, the BlackBerry email client and BlackBerry-to-BlackBerry instant messaging revolutionized the mobile industry.
- When the iPhone debuted in 2007, the world wasn't nearly ready for it. The all-in-one digital music player, camera (2MP!), and Internet-enabled PDA device were introduced by Apple founder Steve Jobs, and the rest is history.
- Devices using the mobile operating system can now be produced by companies like Samsung, LG, HTC, and others thanks to the Android platform.

3.3 KEY BRAND IN MOBILE PHONE INDUSTRY

Manufacturers of mobile phones, providers of operating systems, and carriers are the major stakeholders in the mobile phone market. In the newspaper, the makers are referred to as mobile phone firms. We concentrate on manufacturers of smartphones, which have an operating system akin to that of a computer. These phones, which were invented during the Personal Digital Assistant (PDA) era, have almost all been replaced by smart phones. To be useful to users, a smart phone requires a variety of technologies. These technologies also contain specific software, or "apps," that can be downloaded through app stores in addition to the operating system. Patent battles have been sparked by the highly competitive character of the sector and the demands of high technology. The following is a list of key brands: - Apple - Microsoft - Samsung - HTC - Nokia - Acer - Sony - OPPO - Huawei - ZTE - Motorola - Lava - Blackberry - MaxWest - Lenovo - Micromax - LG - Asus - Kyocera - Vodafone

3.4 WHY WOULD AMAZON WANT TO SELL A MOBILE PHONE?

In 2014 ,Amazon.com entered the mobile phone industry, with the majority of experts speculating that a mystery film indicates that it will introduce a phone with cutting-edge 3-D viewing capabilities.

Amazon has good reason to be interested in the category. The global market for mobile phones is enormous, with close to 2 billion devices supplied each year and more than \$1.6 trillion spent globally on wireless-related services. Mobile devices are becoming more and more important to a variety of businesses as they become the center of the consumer's universe.

The essential activities people attempt to do in their lives change relatively slowly, according to one of the key tenets of Clayton Christensen's famed idea of disruptive innovation. The world develops because inventors find new and improved ways to assist us in achieving the goals we have always tried to achieve, not because our wants, aspirations, or desires change.

Take the significant changes in the music industry. Since the beginning of time as it is known, people have liked listening to music. The largest changes in the industry, however, occurred when creators made it simpler and more convenient for people to listen to the music they want, when they want, and where they want. The first significant musical democratization was brought about by Thomas Edison's phonograph, which made it possible for everyone to enjoy music without needing to pay a live performance, take music lessons, or attend a concert. This trend was accelerated by the ability to hear live sound remotely or a broader range of pre-recorded music thanks to the transmission of sound over the airwaves and reception through a radio.

Floor-standing radios used a lot of power and were relatively expensive. Thus, until Sony made the very portable transistor radio popular in the 1960s, people had a difficult time listening to what they wanted, where they wanted. Teenagers attracted to the device despite its low fidelity in order to listen to late-night baseball games or rock music away from their judgmental parents.

When Sony debuted the Walkman in 1979, it once more made it simpler and easier for people to listen to what they wanted, when they wanted. It's difficult to enjoy music if everyone is shouting transistor radios on the train. The gadget, and its successor the Discman, had one clear drawback: individuals couldn't quickly access their music library while they were away from home. Making mix tapes or carrying around cases filled with numerous CDs let people make up for this.

It became much simpler and easier to listen to the exact music you wanted when and when you wanted thanks to MP3 devices, most notably Apple's iPod. The early iPod advertisements emphasized the benefit of having "1,000 music in your pocket." And lastly, building a music library was even unnecessary thanks to streaming services like Spotify.

Similar principles apply to mobile devices. The first wave of expansion occurred when consumers could easily and increasingly more affordably make phone calls and send text messages while they were on the road thanks to devices from Motorola and Nokia. By making remote e-mail simple, Blackberry let office workers break free from their desks. When Apple and Android-based smart phones put computer-based work and entertainment applications in the palm of your hand, the next wave of growth began.

Aside from the excitement around 3-D technologies, the key question for Amazon as it enters this ostensibly saturated market is if its product makes it simpler or more inexpensive for consumers to engage in activities that have historically been important to them. Experts are dubious, and some have referred to the possible idea as "silly." But allowing customers to examine products before they buy them is one task that a 3-D phone might perform better than current substitutes. People enjoy discovering and acquiring new products, and making the in-store experience available anywhere in the world could make it easier for more people to shop.

The business strategy used by Amazon is arguably even more intriguing. Market disruptions frequently involve the use of a technology that makes things simpler along with an unconventional business strategy. In the current mobile phone business model, service providers finance the phones in exchange for binding customers to two-year service commitments and usage-based fees.

If Amazon's primary goal were to increase retail sales, it might devise entirely different pricing and usage models, subsidize both the phone service and the hardware, perhaps in collaboration with a mobile carrier with a more disruptive business model like T-Mobile, and make money by taking a cut of any transactions made possible by its 3-D platform.

Finally, keep in mind that an innovation's entire influence isn't always immediately obvious when it debuts. It was interesting when Apple introduced the iPod in 2001, but the industry revolutionized when it added the iTunes music store in 2003. Similar to how Google's lightning-fast search technology initially drew people's attention in the late 1990s, the company's AdWords revenue model, which was developed a few years later, is what has made it what it is today.

So in 2014, Amazon has figured out a way to make the difficult simple or the expensive affordable, pay close attention to the business model it intends to use, and, most importantly, people waited to see what the company has planned for the future after the dust settles from the pundit reactions.

3.5 TOPIC MODELLING

Topic modelling is an unsupervised machine learning approach capable of scanning a collection of documents, finding word and phrase patterns within them, and automatically grouping word groups and related expressions that best represent the collection of documents.

A text mining technique called a topic model is used to find obscure subjects in tweets or datasets. Blei, Andrew, and Michael (2003) initially suggested topic modelling techniques utilizing LDA, which are widely employed for many applications, especially for unsupervised analysis. Many specialists from numerous fields have adopted this unsupervised topic modelling technique to help with subject identification in tweets, reviews, etc. Due to LDA's prominence, numerous expansions have been suggested depending on the issue at hand.

The LDA topic model is typically not granular enough, hence Titov and McDonald (2008) proposed a multi grain LDA as a solution. Jo and Oh (2011) also made some intriguing contributions with their sentence-LDA, which searches for hidden themes by connecting words in a sentence to a certain topic.

Supervised topic modeling techniques are frequently utilized as text classifiers (Quercia, 2012), to analyze the ebbs and flows of interest topics across time, and to categorize photographs (Wang and McCallum, 2006). Applications that use topic modeling to support qualitative research and analyze social media material have grown in popularity recently.

3.6 SENTIMENT ANALYSIS:

Sentiment analysis, often known as opinion mining, is a natural language processing (NLP) method for identifying the positivity, negativity, or neutrality of data. Businesses frequently do sentiment analysis on textual data to track the perception of their brands and products in customer reviews and to better understand their target market.

3.7 WHY SENTIMENT ANALYSIS IS IMPORTANT:

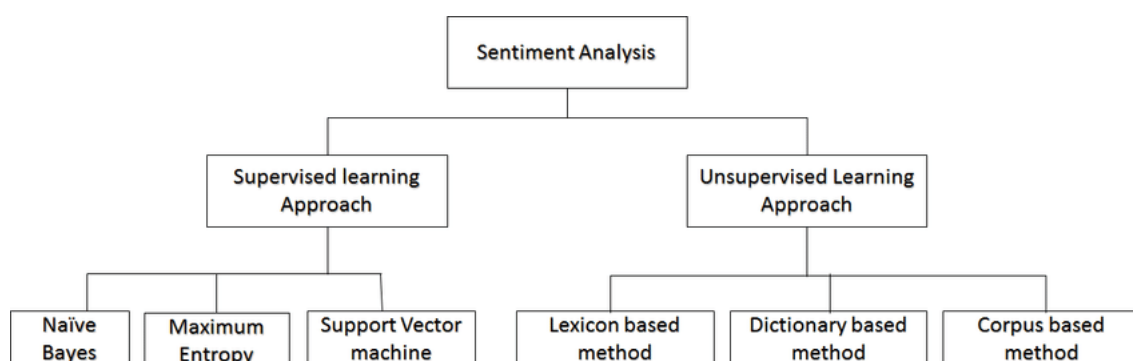
Sentiment analysis is quickly becoming into a crucial tool to monitor and comprehend sentiment in all forms of data because people express their views and feelings more freely than ever before.

Brands can discover what makes customers happy or frustrated by automatically evaluating customer feedback, such as comments in survey replies and social media conversations, in order to modify products and services to suit their needs.

You may learn the reasons why consumers are satisfied or dissatisfied at each point of the customer journey, for instance, by utilizing sentiment analysis to automatically analyze 4,000+ open-ended responses in your customer satisfaction surveys.

Perhaps you want to monitor brand sentiment so that you can identify and address displeased customers as soon as feasible. To determine whether you need to take action, you might attempt to compare sentiment from one quarterly to the next. Then you may delve more deeply into your subjective data to discover the causes of sentiment changes. Sorting Data at Scale, Real-Time Analysis, and Consistent Criteria are a few of the main advantages of sentiment analysis.

3.8 SENTIMENT ANALYSIS APPROACHES



3.9 UNSUPERVISED LEARNING APPROACH

1. **Lexicon based** : Word dictionaries are used in lexicon-based techniques to derive polarity and sentiment. As a result, a wordlist that can be used to convey polarity and output, whether it is positive or negative, is referred to as a sentiment lexicon. The language needs to be extensively developed if sentiment analysis is to be more accurate.
2. **Dictionary based** : Dictionary-based sentiment analysis compares terms in a text or corpus to word lists that have already been constructed as dictionaries. These dictionaries could center on positive/negative words or other concepts like formal/informal language. Each dictionary has a set of words that the user thinks describe a certain trait (such as "positivity" or "impoliteness"). This recipe will concentrate on determining whether a paragraph is positive or negative. Creating or acquiring dictionary files for positive and negative terms, importing the dictionaries and tokenizing them into positive/negative word lists, and importing an analysis text are the first steps.
3. **Corpus based** : Corpus-based approaches imply a data-driven methodology where you will have access to context that you can utilize to your advantage in a machine learning algorithm in addition to sentiment labelling. It is possible to use NLP parsing alone, in combination with rules, or both. The corpus also contains some domain specificity, which might help your algorithm choose the appropriate sentiment label for a word based on its context or domain.

3.10 SUPERVISED LEARNING APPROACH:

The earliest research on sentiment extraction using machine learning was done by (Pang, Lee, and (2002) Vaithyanathan For sentiment analysis, supervised machine learning has been used commonly employed, especially with classifiers like the Support Vector Machine and Naive Bayes (Tanet al., 2009), with a large body of literature contrasting the two to determine which performs better. programs for With the help of features, sentiment analysis has advanced much farther in this strategy such as regulations, word frequency, uni- or n-grams, and others.

3.11 WORD 2 VECTOR:

A method for natural language processing called Word2vec was released in 2013. With the help of a huge text corpus, the word2vec technique employs a neural network model to learn word associations. Once trained, a model like this can identify terms that are similar or suggest new words to complete a sentence. As the name suggests, word2vec uses a specific set of numbers called a vector to represent each unique word. The vectors are carefully selected so that a straightforward mathematical formula (the cosine similarity between the vectors) may be used to determine how similar the words represented by each vector are to one another in terms of meaning.

4. PROBLEM STATEMENT:

The research project reviews Amazon's customers' comments on the mobile phone brands using sentiment analysis and predicts the customer star rating of the particular product using supervised machine learning techniques in machine learning. Given the review data rating label, we will try to get insights about various brands and their ratings using text analytics and build a model to predict overall sentiment. It will be helpful for Amazon to improve its online reputation, increase the sales of the particular product, and increase customer satisfaction. Additionally, to make sure the process is computationally efficient, I created an automation that helps us check the accuracy of all the classification models in a single click. This process will be helpful in terms of computational efficiency and will save the headcount in the industry from not needing to run the entire model again and again.

5. AIM AND OBJECTIVES

5.1 AIM:

By accumulating brand sentiments and automating the model development of coding using the Tkinter library in Python, the research project aims to determine the extent and potential of utilising sentiment analysis on Amazon mobile sale data to quantify a brand's e-reputation. Furthermore, suggest a possible framework for measuring online brand reputation based on the reputation aspects.

5.2 PRIMARY OBJECTIVES

- Examining the data story by charting the visualisation of all the data obtained from the Amazon mobile brand on the Kaggle website
- Determining the most significant variable inside the Amazon data to forecast the rating using sentiment analysis
- Eliminating stopwords from customer reviews to reveal the precise tone of the customer
- Using the Tf-Idf approach to convert the string data into integers in order to model the data.
- Dividing the data using a train-test split based on the most accurate model
- Develop a plan to lower the product's negative perceptions.
- Examine the association between the main themes and the general attitude.
- Automation using Tkinter library to find the best accuracy model and freeze the model as a pickle file for the model deployment

5.3 SECONDARY OBJECTIVES

The research project will also try address the following secondary objectives:

- How should the sentiments be compiled?
- The potential and restrictions of utilising sentiment analysis as a reputation measurement tool in place of more conventional techniques like surveys.

6. METHODOLOGY

For this research project to achieve the set objectives the following applications were used as shown in table :

DESCRIPTION	TYPE
Computer Operating system	Mac OS Big Sur, version 11.5.2
Software	Microsoft Excel 2019
	Jupyter Notebook, version 6.0.1
Programming Languages	Python 3
Data collection platform	Kaggle website

7.TOOL CHOSEN

7.1 PYTHON - Python is an object-oriented, high-level programming language with dynamic semantics that is interpreted. Its high-level built-in data structures, together with dynamic typing and dynamic binding, make it particularly appealing for usage as a scripting or glue language to connect existing components together. Python's concise, easy-to-learn syntax prioritises readability, lowering software maintenance costs. Python has support for modules and packages, which promotes programme modularity and code reuse. The Python interpreter and substantial standard library are free to use and distribute in source or binary form for all major platforms.

Python is frequently embraced by programmers due to the enhanced productivity it delivers. The edit-test-debug cycle is extremely rapid because there is no compilation phase. Debugging Python applications is simple: a bug or incorrect input will never result in a segmentation fault. When the interpreter finds a mistake, it throws an exception. The interpreter produces a stack trace if the programme does not catch the exception. A source level debugger allows you to inspect local and global variables, evaluate arbitrary expressions, create breakpoints, walk through code one line at a time, and more. The debugger is developed in Python, demonstrating the language's introspective

capability. On the other hand, adding a few print statements to the source code is frequently the quickest method to debug a programme: the fast edit-test-debug cycle makes this basic technique quite successful.

7.2 JUSTIFICATION TO USE PYTHON

Python offers several advantages over other machine learning languages. Additionally, Continuous data processing is required for machine learning, and Python's modules allow you to access, handle, and manipulate data. These are some of the most widely used libraries for ML and AI:

- **Scikit-learn** is used to handle fundamental machine learning methods such as clustering, linear and logistic regressions, regression, classification, and others.
- **Pandas** is a high-level data structuring and analysis library. It supports data merging and filtering, as well as obtaining data from other sources such as Excel.
- **Keras** is used for deep learning. It enables quick computations and prototyping by utilising the GPU in addition to the computer's CPU.
- **TensorFlow** is a deep learning framework for creating, training, and deploying artificial neural networks on enormous datasets.
- **Matplotlib** is used to generate 2D plots, histograms, charts, and other types of visualisation.
- Working with computational linguistics, natural language recognition, and processing using **NLTK**.
- Image processing with **Scikit-image**.
- **PyBrain** is a Python library for neural networks, unsupervised and reinforcement learning.
- **Caffe** is a deep learning framework that can alternate between the CPU and the GPU while processing 60+ million photos per day on a single NVIDIA K40 GPU.
- **StatsModels** are used to explore data and statistical procedures.

Additionally, python is very computational efficient in which it is

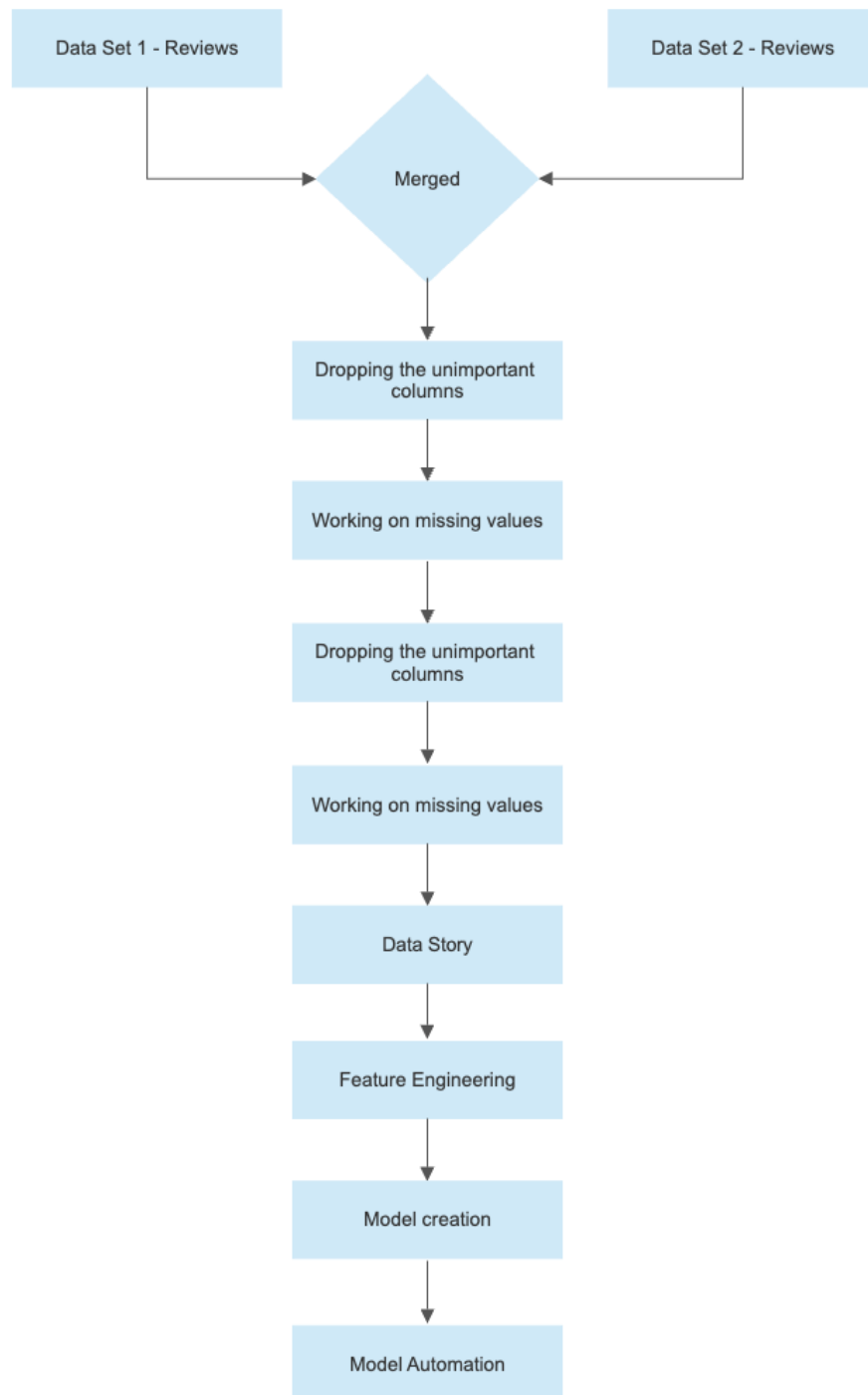
- Easy to learn
- 100% compatible
- Code is clear
- Fast in development
- Libraries are extensive
- Object-oriented
- Open-source and free
- High-level language
- Data structure is built-in

That's why machine learning using python would only be an enormous advantage.

8. RESEARCH PROJECT DESIGN STAGES

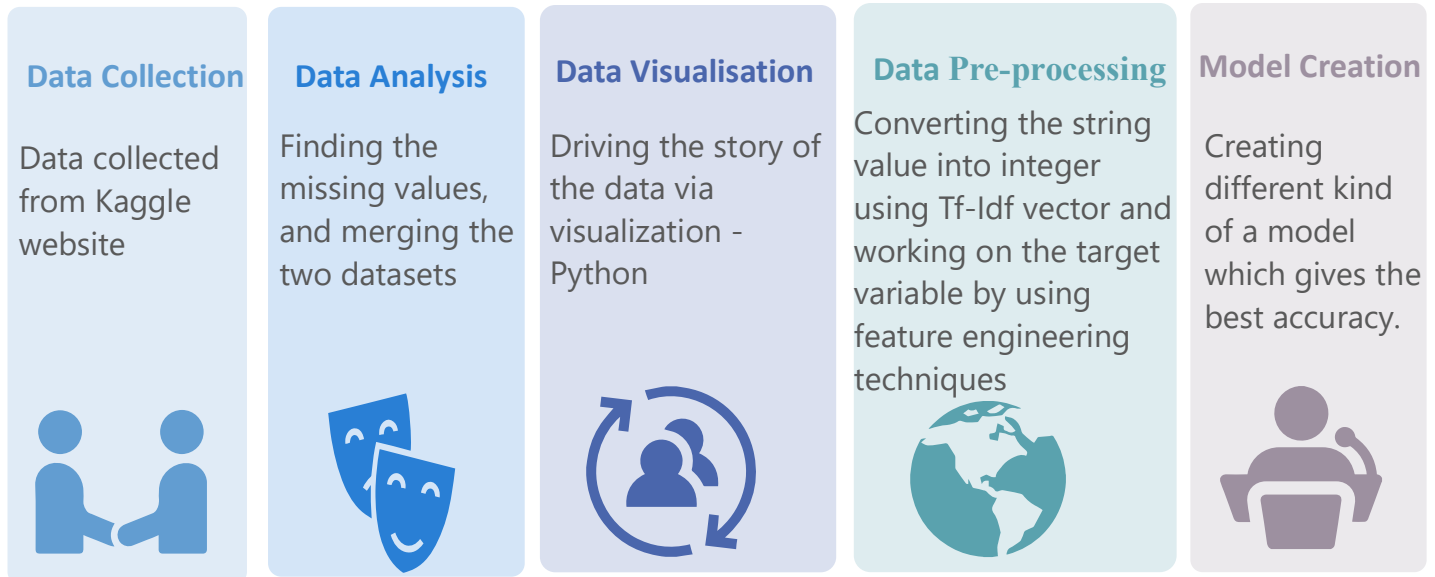
8.1 PROJECT FLOWCHART

The study's step-by-step methodology, including the instruments employed to meet project objectives, is represented by the project flowchart, as illustrated in the image below. The flowchart lists the steps and makes it easier to spot any errors or bottlenecks in the specified process.



8.2 STEPS OF THE PROJECT

The research project was designed into the following stages as shown in figure below



9. PROJECT PROGRAMMING LANGUAGE AND ENVIRONMENT

Table below represents all the tools used as per the project flowchart by each research phase in the project.

RESEARCH STAGE	PACKAGES USED
Data collection	Kaggle Website
Exploratory Data analysis	Microsoft Excel 2019
	Pandas from python
	Numpy from python
Data pre-processing	Sklearn from python
	Tfidfvectorizer from python
	Regex from python
	Nltk from python
	Lambda from python

Data Visualisation	Matplotlib from python Seaborn from python
Model creation	Sklearn from python
Model Automation	Tkinter from python

9.1 DATA COLLECTION:

For my project's initial research phase of data collection, Data collection methods are important because the researcher's methodology and analytical approach influence how the information acquired is used and what explanations it may produce. I used the free Kaggle website to get the necessary information. The data can be found at <https://www.kaggle.com/datasets/tamilarasanpravin/amazon-reviews-and-product-data>. It was uploaded by "Tamilarasan" on the Kaggle website. The dataset has to be large enough to ensure representativeness, relevance, and statistical significance; all of these could be accomplished by gathering data on Kaggle without bias. Another reason for utilising Kaggle as a data gathering platform was its quick access and open availability of data. Another advantage of using the platform was that there was no need to seek for ethical approval because kaggle data is deemed public domain. The two datasets that he posted are available. The two datasets also include product information and reviews from Amazon.

9.1.1 REVIEWS DATASET:

The reviews dataset includes the product, the consumer's feedback, and the customer's name, as well as the star ratings supplied by the customer. The variable and its description are listed in the table below.

Attributes	Description
Asin	Unique item ID of the individual product
Name	Name of the customer who gave the rating and reviews
Rating	The star rating given by the customer
Date	The date of the feedback given by the customer
Verified	Whether the customer is valid or not
Title	The category of the review title
Body	The content of the review
Helpful/votes	Whether the feedbacks are helpful or not

9.1.2 PRODUCT DATASET:

The product dataset contains information on the product's brand, title, URL of the product page, URL of the product picture, URL of the product review, the product's average review, and the total number of reviews provided for the specific product. Each variable is described in detail in the tables below.

Attributes	Description
Asin	Unique item ID of the individual product
Brand	The brand of the particular mobile phone product
Title	The title name of the particular mobile phone product
URL	The exact product URL page to check the product comments
Image	The exact product Image URL page to check the product comments
Rating	The particular product's average star rating
ReviewURL	The particular product review page URL
Total reviews	The particular product's total reviews given by the customer

9.2 EXPLORATORY DATA ANALYSIS :

To clean the data in the second part of the research, **Exploratory Data Analysis**, Exploratory data analysis is critical for every firm. It enables data scientists to assess the data before making any assumptions. It guarantees that the results obtained are valid and related to company objectives and goals. Python was used using the libraries numpy, pandas, and excel. Python was picked because it is a high-level programming language with syntax similar to English. This improves the readability and comprehension of the code. Python is also a highly productive programming language. Python's simplicity allows developers to concentrate on fixing the problem. They do not need to devote a lot of effort to learning the syntax or behaviour of the programming language. You write less code while doing more.

9.3 DATA PRE-PROCESSING :

Data pre-processing is a necessary initial step before applying any machine learning apparatus, because the algorithms learn from the data, and the learning outcome for issue solving is strongly dependent on the right data needed to solve a specific problem - which are referred to as features. Data preprocessing is a necessary job for cleaning the data and preparing it for a machine learning model, which improves the accuracy and efficiency of the machine learning model. we used sklearn library to change the string data from integer. Additionally, we used regex library to find the pattern of the data and replace all the invalid keywords in to null. Furthermore, I used NLTK library to remove stopwords and Lambda function to change the keywords to lowercase. All the above libraries and the basic python is used to predict the sentiment of the product. Furthermore, we can use the basic

python to check the statistical summary to check the mean, median, mode of the data during the pre-processing technique.

9.4 DATA VISUALISATION :

Data visualisation clarifies what information means by providing visual context in the form of maps or graphs. This makes the data more natural for the human mind to understand, making it simpler to see trends, patterns, and outliers in vast data sets. Data visualisation is vital because it allows users to swiftly digest information, increase insights, and make faster decisions. Furthermore, there is a better knowledge of the next measures that must be made to strengthen the company. Finally, a better capacity to keep the audience's attention with material they can grasp. At this stage I have used matplotlib and seaborn library to plot the graph. With the help of visualisation we can share the entire data story and confirm whether the data is bias or not. Additionally, the visualisation can confirm which brand have more negative sentiment and positive sentiment.

9.5 MODEL CREATION:

Machine learning employs two techniques: supervised learning, which involves training a model on known input and output data in order to predict future outputs, and unsupervised learning, which involves discovering hidden patterns or intrinsic structures in input data. A machine learning model is a data file that has been trained to detect specific patterns. You train a model on a collection of data by giving it with an algorithm that it may use to reason about and learn from that data. I used the library sklearn in python to create model. Since it is a classification problem, I used the model Knn, random forest and SVM under supervised learning technique.

9.6 ML AUTOMATION:

The "automatic" aspect of automated machine learning, also known as AutoML, is significant because it enables data architects to collect data and construct algorithms based on past facts. It is present in every sector of our economy, including banking, finance, and insurance. At this stage I have automated the entire model creation part so that we can check all the accuracy of the different models with a single click so that we can confirm the right model and freeze the model as a pickle file and give the model to Machine learning engineer to deploy it. I used the library tkinter from python to automate the entire process.

9.6 HOW MY PROJECT IS DIFFERENT FROM OTHERS:

To transform textual values to numerical values, I utilised the TF-IDF vector. The majority of the papers, however, have not. Furthermore, no model development automation was employed in the review anticipated models.

10. RESEARCH PROJECT STAGES IMPLEMENTATION

10.1 IMPORTING THE DATASET IN JUPYTER NOTEBOOK:

```
df1 = pd.read_csv("reviews.csv")
df2 = pd.read_csv("product_data.csv")
print("Shape of the dataframe of reviews :",df1.shape)
print("Shape of the dataframe of products :",df2.shape)
```

I opened the data sets reviews and product data in a distinct variable and discovered the shape of the data using the pandas library. Pandas offer tremendously simplified data representation. This improves data analysis and comprehension. Data science initiatives with simpler data representation provide better outcomes. The data shape for the opened dataset is,

Shape of the dataframe of reviews : (43932, 8)

Shape of the dataframe of products : (720, 10)

10.2 DATA TYPES:

```
print("data type of reviews \n",df1.dtypes)
print(' '*100)
print("data type of products \n",df2.dtypes)
```

A data type is a property of a piece of data that instructs a computer system how to interpret its value. Understanding data types ensures that data is gathered in the desired format and that each property's value is as expected. To find the data types we used the command `df.dtypes` in python and the solution as follows.

data type of reviews

asin	object
name	object
rating	int64
date	object
verified	bool
title	object

```

body      object
helpfulVotes float64
dtype: object
*****

data type of products
asin      object
brand     object
title     object
url       object
image     object
rating    float64
reviewUrl object
totalReviews int64
price     float64
originalPrice float64
dtype: object

```

We can check that, except for the variables rating, confirmed, and helpful votes, all of the variables in the review dataset are of the object data type. Second, consider the dataset products. Except for the variables rating and total reviews, which are of the float and integer data types, we can establish that all of the variables are of the object data type.

10.3 DROPPING THE UNIMPORTANT COLUMNS:

```

df1 = df1[['asin', 'rating', 'title', 'body']]
df1.head()

```

	asin	rating	title	body
0	B0000SX2UC	3	Def not best, but not worst	I had the Samsung A600 for awhile which is abs...
1	B0000SX2UC	1	Text Messaging Doesn't Work	Due to a software issue between Nokia and Spri...
2	B0000SX2UC	5	Love This Phone	This is a great, reliable phone. I also purcha...
3	B0000SX2UC	3	Love the Phone, BUT...!	I love the phone and all, because I really did...
4	B0000SX2UC	4	Great phone service and options, lousy case!	The phone has been great for every purpose it ...

```

df2 = df2[['asin', 'brand']]
df2.head()

```

	asin	brand
0	B0000SX2UC	NaN
1	B0009N5L7K	Motorola
2	B000SKTZ0S	Motorola
3	B001AO4OUC	Motorola
4	B001DCJAJG	Motorola

To make the model more efficient, we would need to delete unnecessary columns from the dataset, which would help the computer execute the models more quickly. We can observe that the first dataset has just the features Asin, rating, title, and body, whereas the second dataset contains only the features Asin and brand.

10.4 IDENTIFICATION AND TREATMENT OF ANY MISSING VALUES:

When dealing with real-world datasets, such as those available on Kaggle, missing values are prevalent. Missing data can be caused by a human component (for example, a person purposefully neglecting to reply to a survey question), an issue with electrical sensors, or other circumstances.

Missing data (or missing values) are data values for variables that are not saved in the observation of interest. Missing data is a reasonably prevalent problem in practically any research, and it can have a substantial impact on the conclusions that can be taken from the data.

```
def missing_values(x):  
    print(x.isnull().sum())  
missing_values(df1)  
missing_values(df2)
```

Solution:

asin	0
rating	0
title	5
body	12

asin	0
brand	4

The solution clearly says that the features title and body have 5 and 12 missing values in the first dataset, and the brand feature has 4 missing values in the second dataset. Since the missing values are literally low in percentage of the data, I have removed the entire missing values data from the dataset.

10.5 DATA MERGING:

The merge() function merges two DataFrames together and modifies their information using the provided method. To merge both the dataset, I used the pandas function pd.merge() to merge the data so that we can process it further. The below image clearly shows the merged data.


```
df = pd.merge(df1,df2, on = 'asin')
df.head()
```

	asin	rating	title		body	brand
0	B0009N5L7K	1	Stupid phone		DON'T BUY OUT OF SERVICE	Motorola
1	B0009N5L7K	4	Exellent Service	I have been with nextel for nearly a year now ...		Motorola
2	B0009N5L7K	5	I love it	I just got it and have to say its easy to use,...		Motorola
3	B0009N5L7K	1	Phones locked	1 star because the phones locked so I have to ...		Motorola
4	B0009N5L7K	5	Excellent product	The product has been very good. I had used thi...		Motorola

10.6 STATISTICAL SUMMARY:

The describe() method generates a summary of statistics for the columns in the DataFrame. This function returns the mean, standard deviation, and interquartile range (IQR) values. Furthermore, the method eliminates character columns and provides a summary of numeric columns.

```
def statistical_summary(df):
    print(df.describe())
statistical_summary(df)
```

```

              rating
count  43839.000000
mean      3.677456
std       1.630336
min       1.000000
25%       2.000000
50%       5.000000
75%       5.000000
max       5.000000
```

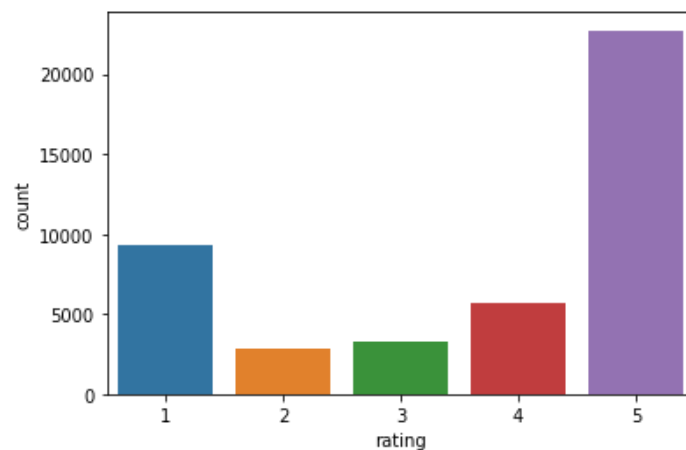
According to the given answer, there are 43829 rows in the dataset. Furthermore, the only integer type attribute that we can validate has a minimum value of 1 and a maximum value of 5 is rating. This implies that the consumer provided comments on a scale of 1 to 5, with a mean value of 3.6 and a standard deviation of 1.63.

10.7 EXPLORATORY DATA ANALYSIS

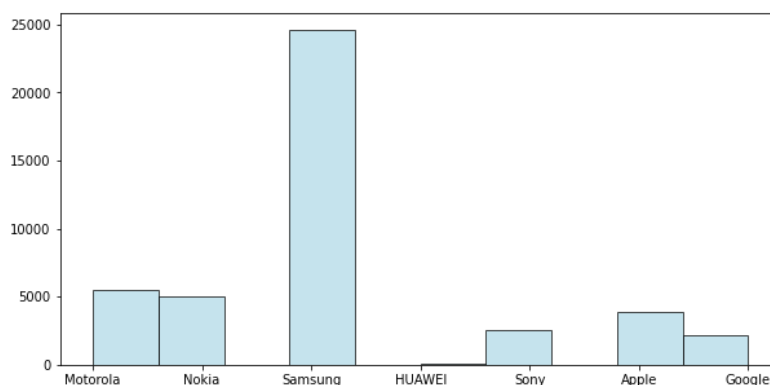
The process of developing maps and other visuals when dealing with somewhat unknown geographic data is known as exploratory visualisation. These maps often serve a particular purpose and serve as an aid in the expert's endeavour to solve a (geo) issue.

To further understand the dataset, an exploratory data analysis was performed using Python's seaborn and matplotlib modules to summarise the major aspects of the data. This stage of the research is critical because it helps with getting a sense of what's inside the data set, translating it into a visual medium to quickly identify its features, such as interesting curves, lines, trends, or anomalous outliers, and it's used to find any interesting storylines in the data.

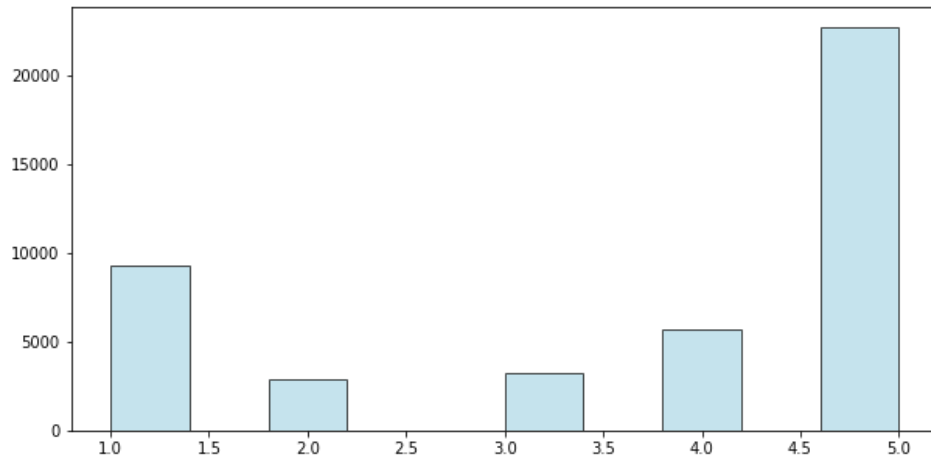
We previously knew from the dataset that customers were rated from 1 to 5. Furthermore, we discovered from the graph below that the majority of consumers (over 25000) rated Amazon mobile phones as 5 stars. Furthermore, the next highest rating was given as 1, which is near to 10000 and should be observed since items with lower ratings will not be marketed in the future. We used count plot for the below graph. A count plot is analogous to a histogram over a category variable rather than a quantitative variable. The fundamental API and settings are the same as for barplot(), allowing you to compare counts across nested variables.



The graph below depicts the amount of consumers who were given ratings for various brands. We can certify that the number of consumers who rated the brand "Samsung" is quite close to 25000. In comparison, the brand "HUAWEI" had the fewest evaluations, which were close to 0 -100. Finally, additional brands have numbers ranging from 2500 to 5000. We used distribution plot for the below graph. Distribution plots evaluate the distribution of sample data graphically by comparing the empirical distribution of the data to the theoretical values anticipated given a specific distribution.



The below distribution plot is for the feature rating in which we can identify how the ratings are distributed. We can confirm that the rating 5 is distributed well among all the ratings.



10.8 ASSIGNING THE TARGET VARIABLE

The target variable is the one whose values are anticipated and modelled by other variables. A predictor variable is one whose values are used to forecast the value of the target variable.

The target variable is the aspect of a dataset that you wish to better understand. It is the variable that the user wants to forecast using the remaining data. In most cases, the target variable is determined using a supervised machine learning method. An algorithm of this type learns patterns from past data and discovers correlations between other sections of your dataset and the goal. Depending on the aim and available data, target variables may differ.

10.9 WHY TARGET VARIABLE IS IMPORTANT ?

When developing a machine learning system for evaluating customer attrition in the telecommunications business, you must spend weeks or even months studying how some consumers unsubscribe and others renew. You can start utilising machine learning in production after you have enough training instances to develop an accurate machine learning model.

```
def assigning_values(x):  
    if x > 3:  
        return 1  
    elif x < 3:  
        return -1  
    elif x == 3:  
        return 0  
df['Sentiment'] = df['rating'].map(assigning_values)  
df['Sentiment'].head()
```

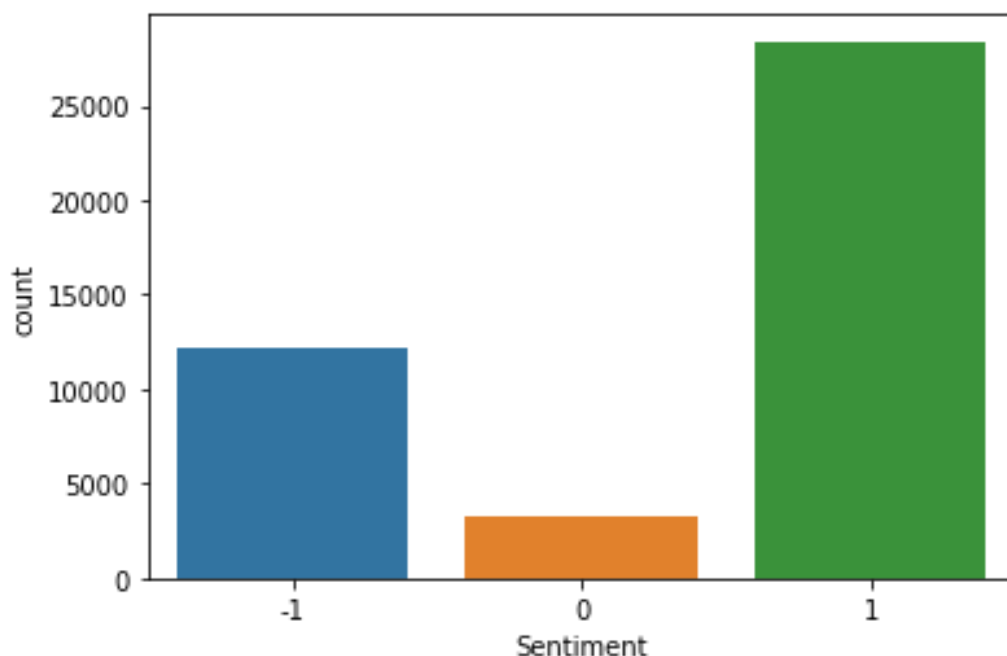
With the use of a Python "if" statement, I assigned all ratings over 3 as 1, ratings less than 3 as -1, and ratings equal to 3 as 0. I portray negative feelings as -1, neutral sentiments as 0 and good sentiments as 1. Additionally, I used the "map" statement to update all of the assigned values in the target column.

10.10 TARGET VARIABLE ANALYSIS

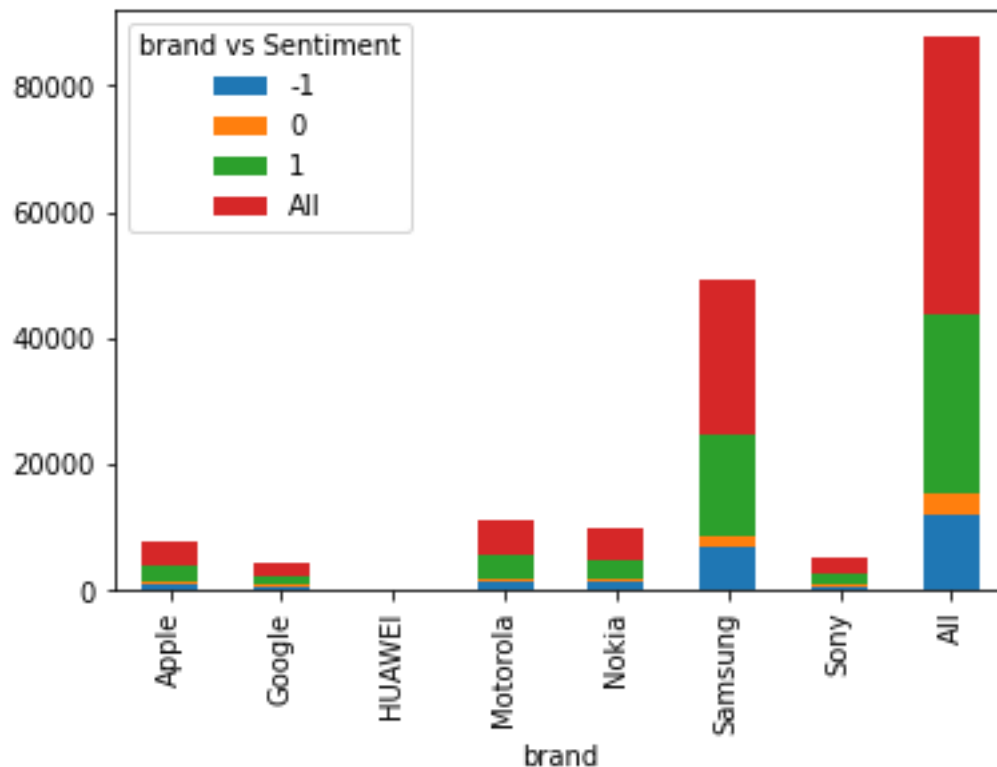
In general, the target variable should have a very uniform distribution, with as near to a 50/50 split as feasible in the binary situation. If the variable is skewed to one side or the other, the model will have a more difficult time evaluating the other predictor variables. Consider oversampling your data if your distribution is unequal.

The target variable of a dataset is the aspect of a dataset about which you wish to learn more. A supervised machine learning algorithm learns patterns from previous data and discovers links between various properties of your dataset and the goal.

The target variable will differ based on the business aim and the data provided. Assume you wish to utilise sentiment analysis to determine if tweets about your company's brand are favourable or bad. Word tokens, parts of speech, and emoticons are just a few examples of twitter characteristics. A model cannot learn how those characteristics relate to sentiment unless it is first shown samples of good and negative tweets (the target).



According to the graph, the positive sentiment is more prevalent. As a result, there is a potential that the model will become overfit. However, we cannot upscale or downscale the data since the client feedback in this circumstance is likely to be skewed.

Brand vs sentiment:

The graph above plainly shows that the brand Samsung received a large number of favourable, negative, and neutral responses. Other brands do have very little amount of all the sentiments. The "HUAWEI" does not have the most remarks. Hence, we can concentrate more on the brand Samsung.

10.11 LOWERCASING THE STRING DATA

```
new_df= new_df.apply(lambda x: x.astype(str).str.lower())
new_df.head()
```

	title	body
0	stupid phone	don't buy out of service
1	exellent service	i have been with nextel for nearly a year now ...
2	i love it	i just got it and have to say its easy to use,...
3	phones locked	1 star because the phones locked so i have to ...
4	excellent product	the product has been very good. i had used thi...

Lower casing is the process of converting a word to lower case (NLP -> nlp). Words like Book and book denote the same thing, however in the vector space model, they are represented as two separate words since they are not changed to lower case (resulting in more dimensions). As a result, I utilised Python's lambda function to transform all of the string data to lowercase. The result is seen in the figure above.

10.12 REMOVING THE PUNCTUATION:

Because punctuation marks are frequently used in text, removing them is an important NLP pre-processing step. These marks - used to divide text into sentences, paragraphs, and phrases - affect the results of any text processing approach, particularly those that rely on the occurrence frequencies of words and phrases. Stop-words, which are often used in language, are deleted before any NLP procedure. Stop-words are words that are regularly used without any extra information, such as articles, determiners, and prepositions. By deleting these frequently used terms from the text, we may instead concentrate on the crucial words. Many studies have been offered in terms of: studying the influence of stop-word removal pre-processes, such as Silva et al.12, and giving techniques for listing stop-words, such as Klatt et al.13.

```
import re
new_df['title'] = new_df['title'].apply(lambda x: re.findall("[\w"]+", x))
new_df['body'] = new_df['body'].apply(lambda x: re.findall("[\w"]+", x))
```

As a consequence, we used Python's lambda function to find all of the punctuation in the variable title and body and replaced it with null. The results are also shown in the image below.

```
new_df.head()
```

	title	body
0	[stupid, phone]	[don't, buy, out, of, service]
1	[exellent, service]	[i, have, been, with, nextel, for, nearly, a, ...
2	[i, love, it]	[i, just, got, it, and, have, to, say, its, ea...
3	[phones, locked]	[1, star, because, the, phones, locked, so, i,...
4	[excellent, product]	[the, product, has, been, very, good, i, had, ...

10.13 REMOVING STOP WORDS

Stop words are often used terms such as 'if,' 'but,' 'we,' 'he,' 'she,' and 'they.' We can generally eliminate these terms without affecting the semantics of the text, and doing so frequently (but not always) enhances a model's performance. When we start utilising longer word sequences as model features, removing these stop words becomes much more beneficial.

```
def remove_stopwords(x):
    return [w.lower() for w in x if not w.lower() in stop_words]

new_df['title'] = new_df['title'].map(remove_stopwords)
new_df['body'] = new_df['body'].map(remove_stopwords)
new_df.head()
```

As a result, we can see that the below output shows the string data.

	title	body
0	[stupid, phone]	[buy, service]
1	[excellent, service]	[nextel, nearly, year, started, time, last, ye...
2	[love]	[got, say, easy, use, hear, person, talking, f...
3	[phones, locked]	[1, star, phones, locked, pay, additional, fee...
4	[excellent, product]	[product, good, used, cell, phone, one, projec...

We can observe that the string values are separated by commas after we remove the stop words. However, we must remove the comma and allow the string values to emerge from the list data type. To get rid of it, I used the same lambda function to combine the data for the variable body and delete the string values from the list, as seen in the figure below.

```
new_df['body'] = new_df['body'].apply(lambda x: ' '.join(x))
new_df.head()
```

	title	body
0	[stupid, phone]	buy service
1	[excellent, service]	nextel nearly year started time last year moto...
2	[love]	got say easy use hear person talking fine prob...
3	[phones, locked]	1 star phones locked pay additional fees unlock
4	[excellent, product]	product good used cell phone one projects work...

10. 14 TRANSFORMING “BODY” USING TF-IDF VECTOR

A text vectorizer that converts text into a useful vector is term frequency-inverse document frequency. It combines two ideas: term frequency (TF) and document frequency (DF) (DF).

The term frequency is the number of times a certain phrase appears in a document. The frequency of occurrence of a phrase in a document reflects its importance. Term frequency depicts each text in the data as a matrix, with rows representing the number of documents and columns representing the number of different words across all documents.

The amount of papers that include a certain phrase is referred to as document frequency. The term's document frequency reflects how prevalent it is.

The weight of a term is inverse document frequency (IDF), which seeks to minimise the weight of a phrase if its occurrences are spread across all documents. IDF is computed as follows:

$$idf_i = \log\left(\frac{n}{df_i}\right)$$

Where idf_i represents the IDF score for term i , df_i represents the number of documents that include term i and n is the total number of documents. The lower the IDF for a term, the greater the DF of the term. When the number of DF is equal to n , which indicates that the term appears in all documents, the IDF is 0 since $\log(1)$ is zero; if in doubt, simply add this term to the stopwords list because it doesn't give much information.

The TF-IDF score is just a multiplication of the term frequency matrix by its IDF, and it may be calculated as follows:

$$w_{i,j} = tf_{i,j} \times idf_i$$

```
from sklearn.feature_extraction.text import TfidfVectorizer
# tf_idf = TfidfVectorizer(ngram_range=(1,2))
vectorizer = TfidfVectorizer(max_features=100)
X = vectorizer.fit_transform(new_df['body'])
X.todense().shape
```

From the image above, we observe that with the help of the library `TfidfVectorizer` from `SKlearn`, we changed all the string values into integers. As a result, I set `max_features=100`, and you're ready to go. If `max_features` is set to `None`, the TF-IDF transformation takes into account the whole corpus. Otherwise, passing 5 to `max_features` would mean constructing a feature matrix from the top 5 most often occurring terms in text documents. The converted table is shown below.


```
x = pd.DataFrame(X.todense(), columns=vectorizer.get_feature_names())
x
```

/usr/local/lib/python3.7/dist-packages/sklearn/utils/deprecation.py:87: FutureWarning: Function get_feature_names is deprecated; get_feature_names
warnings.warn(msg, category=FutureWarning)

	also	amazon	android	another	app	apps	back	battery	best	better	...	verizon	want	way	well	windows	without	work
0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	...	0.000000	0.000000	0.0	0.000000	0.0	0.000000	0.000000
1	0.000000	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.683445	0.0	...	0.000000	0.000000	0.0	0.000000	0.0	0.000000	0.000000
2	0.000000	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	...	0.000000	0.000000	0.0	0.000000	0.0	0.000000	0.000000
3	0.000000	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	...	0.000000	0.000000	0.0	0.000000	0.0	0.000000	0.000000
4	0.000000	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	...	0.000000	0.000000	0.0	0.000000	0.0	0.000000	0.000000
...
43834	0.000000	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	...	0.000000	0.000000	0.0	0.000000	0.0	0.000000	0.000000
43835	0.000000	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	...	0.000000	0.000000	0.0	0.000000	0.0	0.000000	0.239405
43836	0.000000	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	...	0.000000	0.000000	0.0	0.283120	0.0	0.000000	0.273391
43837	0.194963	0.0	0.142352	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	...	0.275654	0.072465	0.0	0.124634	0.0	0.074142	0.000000
43838	0.000000	0.0	0.000000	0.0	0.0	0.0	0.542567	0.0	0.000000	0.0	...	0.000000	0.000000	0.0	0.000000	0.0	0.000000	0.000000

Finally, we have saved the target variable in a variable called Y. Except the target variable, we saved the other variable as X to split the data.

10.15 TRAIN TEST SPLIT

The train-test split is a strategy for assessing a machine learning algorithm's performance. It may be used for any supervised learning technique and can be utilised for classification or regression tasks.

The process includes partitioning a dataset into two subgroups. The first subset, known as the training dataset, is utilised to fit the model. The second subset is not used to train the model; instead, the model is fed the dataset's input element, and predictions are generated and compared to predicted values. The second dataset is known as the test dataset.

- **Train Dataset:** This dataset is used to fit the machine learning model.
- **Test Dataset:** Used to assess the fit of the machine learning model.

The goal is to assess the machine learning model's performance on new data that was not used to train the model.

This is how we anticipate using the model in practise. To put it another way, we want to fit it to existing data with known inputs and outputs and then make predictions on fresh cases in the future where we don't have the expected output or goal values. When a sufficiently big dataset is available, the train-test technique is appropriate.

10.15.1 WHEN TO USE TRAIN-TEST SPLIT

Each predictive modelling task has its own definition of "sufficiently large." It signifies that there is enough data to divide the dataset into train and test datasets, and that each train and test dataset is a good representation of the problem domain. This necessitates that the original dataset also be a good representation of the issue domain.

A appropriate representation of the problem domain means that there are enough entries to cover all of the domain's frequent and rare occurrences. This might refer to input variable combinations encountered in practise. Thousands, hundreds of thousands, or millions of instances may be required.

When the available dataset is small, the train-test process is ineffective. The reason for this is because when the dataset is divided into train and test sets, the training dataset will not include enough data for the model to learn an efficient mapping of inputs to outputs. There will also be insufficient data in the test set to evaluate the model's performance appropriately. The predicted performance may be too optimistic (excellent) or extremely pessimistic (poor) (bad).

If you don't have enough data, the k-fold cross-validation approach is a good alternative model assessment method. Another reason to employ the train-test split assessment process, in addition to dataset quantity, is computational efficiency.

Some models are extremely expensive to train, making frequent assessment, as employed in other techniques, impossible. Deep neural network models are one example. The train-test approach is widely employed in this instance. Alternatively, a project may have an efficient model and a large dataset but need a rapid assessment of model performance. In this case, the train-test split technique is used once more.

Using random selection, samples from the original training dataset are divided into two groups. This is done to guarantee that the train and test datasets reflect the original dataset.

10.15.2 HOW TO CONFIGURE THE TRAIN-TEST SPLIT

The size of the train and test sets is the procedure's key configurable parameter. For either the train or test datasets, this is most typically given as a percentage between 0 and 1. A training set with a size of 0.67 (67%), for example, indicates that the leftover percentage of 0.33 (33%) is assigned to the test set.

There is no such thing as an appropriate split %. You must select a split % that matches the objectives of your project, taking into account factors such as:

- Computational cost in training the model.
- Cost of computation in assessing the model.
- Representativeness of the training set
- Representativeness of the test set

However, popular split percentages include:

- 80% for training, 20% for testing
- 67% train, 33% test

- 50% training, 50% testing

Now that we've covered the train-test split model assessment technique, let's look at how we might apply it in Python.

```
x_train, x_test, y_train, y_test = train_test_split(X, Y, test_size = 0.3 )  
  
x_train.shape, x_test.shape, y_train.shape, y_test.shape  
(30687, 100), (13152, 100), (30687,), (13152,))
```

As per the above image, I split the data with 0.3 as test size which means 70% of the data goes to training set and 30% of the data goes to testing set. Additionally, the shape of the training set is 30687 and test set is 13152.

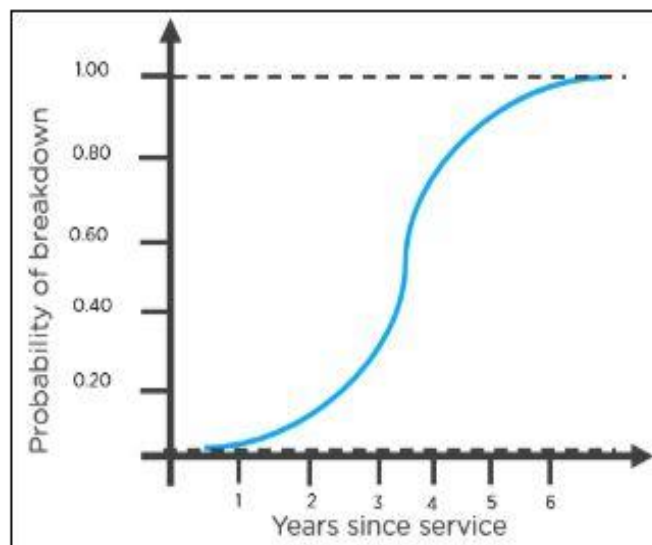
10.16 MODEL CREATION:

10.16.1 LOGISTIC REGRESSION:

Logistic regression is a statistical approach for developing machine learning models using dichotomous dependent variables, i.e. binary. Logistic regression is a statistical technique used to describe data and the connection between one dependent variable and one or more independent variables. Independent variables may be nominal, ordinal, or interval in nature.

The term "logistic regression" is derived from the logistic function that it employs. The sigmoid function is another name for the logistic function. This logistic function has a value between zero and one.

The logistic function shown below may be used to calculate the likelihood of a car breaking down based on how long it has been since it was serviced.



Advantages of Logistic regression:

- When the data is linearly separable, logistic regression performs better.
- Because it is highly interpretable, it does not need a large number of computer resources.
- Scaling the input characteristics is simple—no adjustment is required.
- The logistic regression model is simple to construct and train.
- It provides a measure of a predictor's relevance (coefficient size) as well as the direction of relationship (positive or negative)

Multinomial logistic regression is a straightforward expansion of binary logistic regression that allows for more than two dependent or outcome variable categories. Multinomial logistic regression, like binary logistic regression, use maximum likelihood estimation to assess the likelihood of category membership. However, we have 3 target features available as a sentiment. Hence, we used multiple logistic regression in this situation.

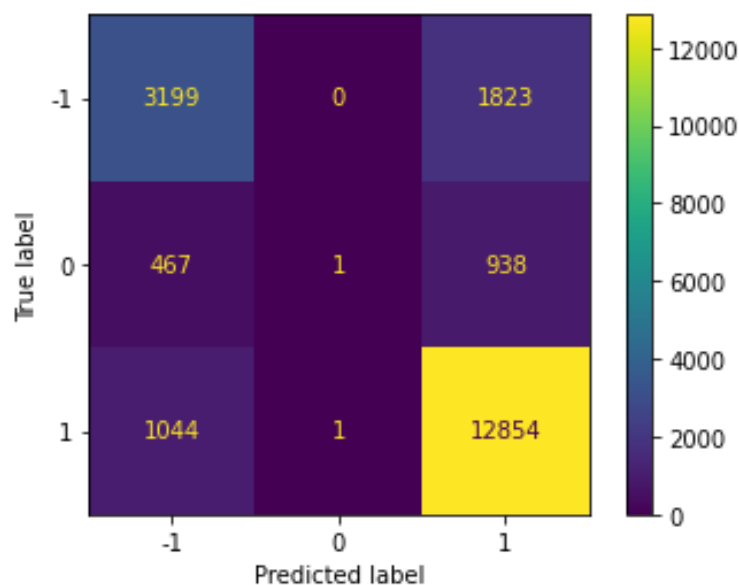
```

classifier = LogisticRegression(random_state = 0)
classifier.fit(x_train, y_train)
y_pred_lr = classifier.predict(x_test)
LR_score = accuracy_score(y_test, y_pred_lr)
print("Accuracy score (LR): ", LR_score)

LR_accuracy_matrix = confusion_matrix(y_test, y_pred_lr)
sns.heatmap(pd.DataFrame(LR_accuracy_matrix), annot = True, cmap="YlGnBu" ,fmt='g').set_title('Voting_Classifier')

```

The accuracy for the logistic regression is 0.78 which means the model gives 78% of accurate results of the data.

Confusion Matrix

Confusion matrix It is a table used in classification issues to determine where mistakes in the model occurred. The rows indicate the actual classes for which the results should have been calculated. Whereas the columns indicate our expectations. This table makes it simple to identify whose forecasts are incorrect.

From the above table, we can say that

- Correct Detection of Negative sentiment (-1) is 3199 out of 5022 data
- Correct Detection of Positive sentiment (1) is 12854 out of 13890 data
- Correct Detection of Neutral sentiment (0) is 1 out of 1405 data

Classification report:

A classification report is a machine learning performance evaluation statistic. It is used to display your trained classification model's accuracy, recall, F1 Score, and support.

Metrics	Definition
Precision	Precision is defined as the ratio of true positives to the sum of true and false positives.
Recall	Recall is defined as the ratio of true positives to the sum of true positives and false negatives.
F1 Score	The F1 is the weighted harmonic mean of precision and recall. The closer the value of the F1 score is to 1.0, the better the expected performance of the model is.
Support	Support is the number of actual occurrences of the class in the dataset. It doesn't vary between models, it just diagnoses the performance evaluation process.

Target	Precision	Recall	F1-score	Support
-1	0.67	0.64	0.65	5057
0	0.4	0	0	1435
1	0.82	0.92	0.87	13835
accuracy			0.79	20327
macro avg	0.63	0.52	0.51	20327
weighted avg	0.75	0.79	0.75	20327

According to the logistic regression model, the neutral sentiment is meaningless because the recall and F1 score are both 0 for the specific category, indicating that the model does not accurately predict the neutral sentiment. However, because the data in the negative sentiment category is limited, the accuracy, recall, and f1 score are 0.67, 0.64, and 0.65, respectively. If we collect additional data, we can enhance the accuracy and create a better model.

10.16.2 NAÏVE BAYE'S CLASSIFIER

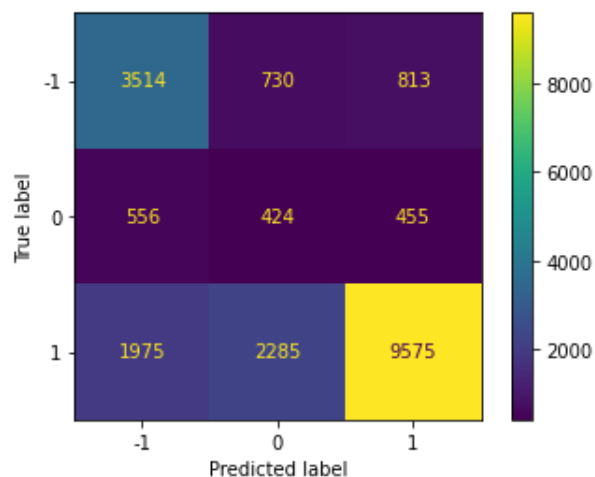
Naive Bayes is a probabilistic classifier that returns the likelihood of a test point belonging to a class rather than its label. It is one of the most fundamental Bayesian network models, but when paired with kernel density estimation, it may achieve higher levels of accuracy. Unlike many other ML algorithms, which can often handle both Regression and Classification tasks, this algorithm is exclusively appropriate for Classification tasks.

Naive The Bayes method is regarded as naïve since the assumptions it makes are nearly hard to detect in real-world data. It employs conditional probability to compute a product of component probabilities. This indicates that, given the class variable, the algorithm assumes the existence or absence of a specific characteristic of a class that is unrelated to the presence or absence of any other feature (absolute independence of features).

The Bayes theorem (sometimes known as the Bayes rule) is based on conditional probability. In conditional probability, the occurrence of one event is conditional on the occurrence of another. The Bayes theorem argues that given two occurrences A and B,

$$P(A|B)=P(A\cap B)/P(B)=P(A)\cdot P(B|A)/P(B)$$

The accuracy for the logistic regression is 0.66 which means the model gives 66% of accurate results of the data.



Target	Precision	Recall	F1-score	Support
-1	0.58	0.69	0.63	5057
0	0.12	0.30	0.17	1435
1	0.88	0.69	0.78	13835
accuracy			0.66	20327
macro avg	0.53	0.56	0.53	20327
weighted avg	0.75	0.66	0.70	20327

From the above table, we can say that

- Correct Detection of Negative sentiment (-1) is 3514 out of 5022 data
- Correct Detection of Positive sentiment (1) is 9575 out of 13890 data
- Correct Detection of Neutral sentiment (0) is 424 out of 1405 data

The accuracy is 66% according to the Nave Bayes model. However, the data is skewed in order to achieve greater accuracy. It is overfitted to the target variable positive sentiment since its value counts are larger than the other two target features.

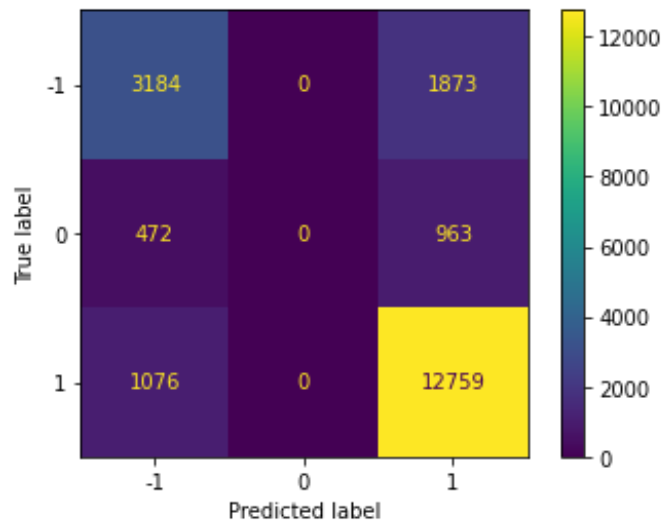
10.16.3 SUPPORT VECTOR MACHINE

SVM is a supervised machine learning technique that may be used for both classification and regression. Though we call them regression issues, they are best suited for categorization. The SVM algorithm's goal is to identify a hyperplane in an N-dimensional space that clearly classifies the input points.

The following are the benefits of support vector machines:

- Effective in high-dimensional environments.
- When the number of dimensions exceeds the number of samples, the method remains effective.
- It also saves memory by using a subset of training points in the decision function (called support vectors).
- The decision function can be provided using several Kernel functions. Common kernels are given, however custom kernels can also be specified.

The accuracy for the logistic regression is 0.78 which means the model gives 66% of accurate results of the data.



Target	Precision	Recall	F1-score	Support
-1	0.67	0.63	0.65	5057
0	0.00	0.00	0.00	1435
1	0.82	0.92	0.87	13835
accuracy			0.78	20327
macro avg	0.50	0.52	0.51	20327
weighted avg	0.72	0.78	0.75	20327

From the above table, we can say that

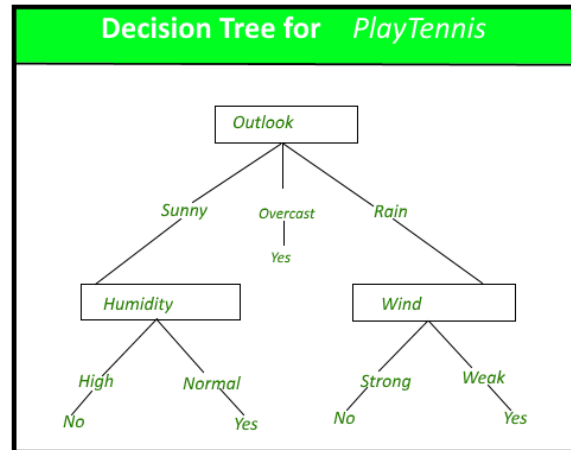
- Correct Detection of Negative sentiment (-1) is 3184 out of 5022 data
- Correct Detection of Positive sentiment (1) is 12759 out of 13890 data
- Correct Detection of Neutral sentiment (0) is 0 out of 1405 data

From the SVM model, we can say that the accuracy is 78%. Moreover, this model's performance was the same as the logistic regression model's performance. However, it did not find the neutral sentiment properly because the neutral sentiment value is literally low in count. Maybe in the future, if we upscale or downscale the data, we might get better accuracy in the model. However, for industrial purposes, 78% is better accuracy.

10.16. 4 DECISION TREE CLASSIFIER:

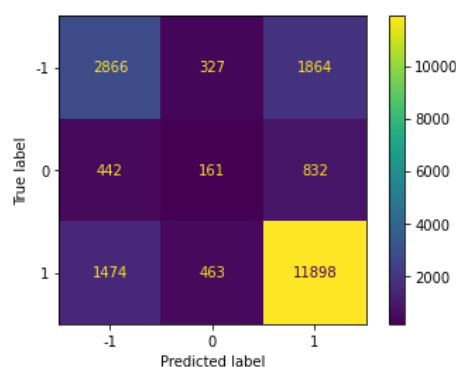
The most powerful and widely used tool for categorization and prediction is the Decision Tree. A Decision tree is a tree structure that looks like a flowchart, with each internal node representing a test

on an attribute, each branch representing a test outcome, and each leaf node (terminal node) holding a class label.



Decision Tree Construction: A tree may be "learned" by dividing the source set into subgroups based on an attribute value test. This method is performed recursively on each derived subset, which is known as recursive partitioning. When the subset at a node all has the same value of the target variable, or when splitting no longer adds value to the predictions, the recursion is finished. Because the development of a decision tree classifier requires no domain expertise or parameter setup, it is suitable for exploratory knowledge discovery. High-dimensional data may be handled via decision trees. The decision tree classifier is often accurate. Decision tree induction is a common inductive way of learning classification information.

Gini Index: The Gini Index is a statistic that determines how accurate a classification is. The Gini index assigns a score between 0 and 1, with 0 representing all data belonging to one class and 1 representing a random distribution of items within classes. In this situation, we wish to have the lowest Gini index score possible. The Gini Index will be used as an assessment tool to assess our Decision Tree Model.



Target	Precision	Recall	F1-score	Support
-1	0.60	0.57	0.58	5057
0	0.17	0.11	0.13	1435
1	0.82	0.86	0.84	13835
accuracy			0.73	20327
macro avg	0.53	0.51	0.52	20327
weighted avg	0.72	0.73	0.72	20327

From the above table, we can say that

- Correct Detection of Negative sentiment (-1) is 2866 out of 5022 data
- Correct Detection of Positive sentiment (1) is 11898 out of 13890 data
- Correct Detection of Neutral sentiment (0) is 161 out of 1405 data

Additionally, the accuracy of the model is 73%.

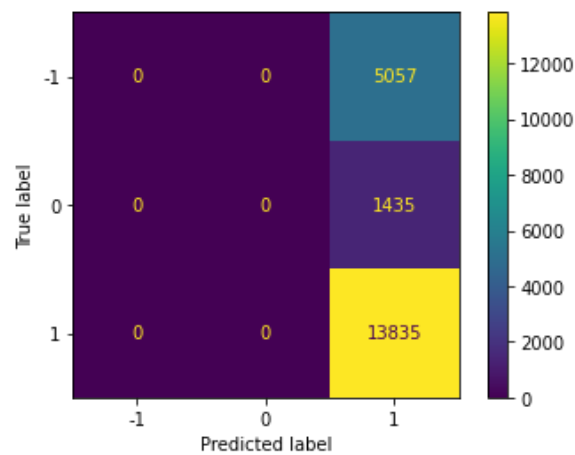
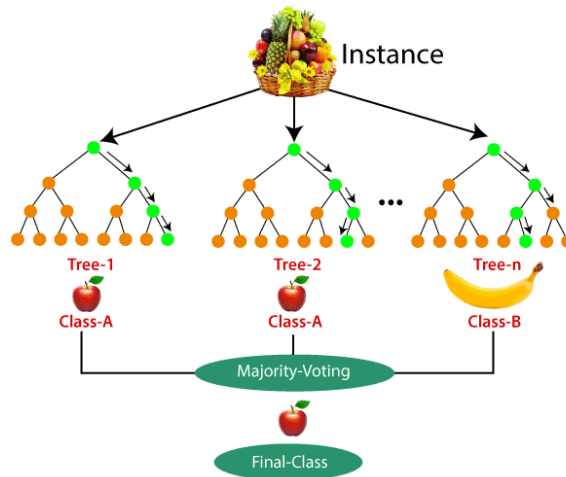
10.16.5 RANDOM FOREST CLASSIFIER:

Random Forest is a well-known machine learning algorithm from the supervised learning approach. It can be used for both Classification and Regression problems in ML. It is built on the notion of ensemble learning, which is a method that involves integrating several classifiers to solve a complicated issue and enhance the model's performance.

"Random Forest is a classifier that comprises a number of decision trees on various subsets of the provided dataset and takes the average to enhance the predicted accuracy of that dataset," as the name implies. Instead than depending on a single decision tree, the random forest collects the forecasts from each tree and predicts the final output based on the majority vote of predictions.

10. 16.5.1 WHY USE OF RANDOM FOREST

- It requires shorter training time than other algorithms.
- It predicts output with great accuracy, and it works efficiently even on big datasets.
- It can also retain accuracy when a significant amount of data is absent.



Target	Precision	Recall	F1-score	Support
-1	0.00	0.00	0.00	5057
0	0.00	0.00	0.00	1435
1	0.68	1.00	0.81	13835
accuracy			0.68	20327
macro avg	0.23	0.33	0.27	20327
weighted avg	0.46	0.68	0.55	20327

From the above table, we can say that

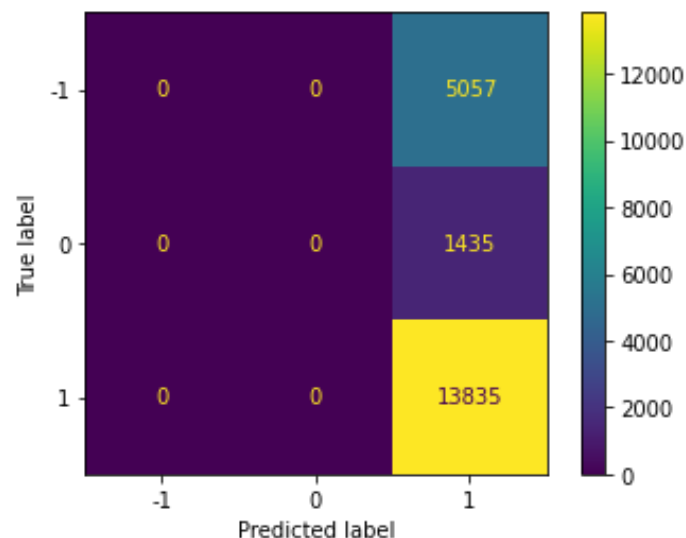
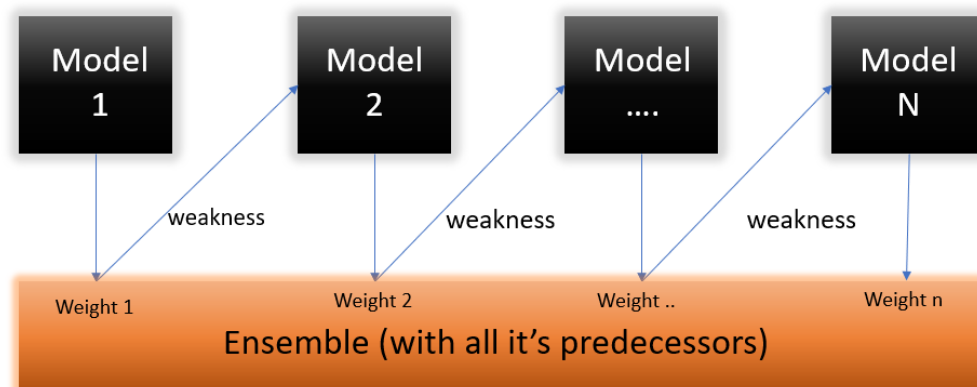
- Correct Detection of Negative sentiment (-1) is 0 out of 5022 data
- Correct Detection of Positive sentiment (1) is 13835 out of 13890 data
- Correct Detection of Neutral sentiment (0) is 0 out of 1405 data

Furthermore, the Random Forest model has a 68% accuracy. However, the model failed to accurately predict negative and neutral mood. To improve the accuracy of the same model, we must experiment with various hyperparameters.

10.16.6 ADA BOOSTING

AdaBoost, also known as Adaptive Boosting, is a Machine Learning approach that is utilised as an Ensemble Method. The most frequent AdaBoost method is decision trees with one level, which is decision trees with just one split. These trees are often referred to as Decision Stumps.

This approach constructs a model and assigns equal weights to all data points. It then applies larger weights to incorrectly categorised points. In the following model, all points with greater weights are given more weight. It will continue to train models till a low error is returned.



Target	Precision	Recall	F1-score	Support
-1	0.66	0.59	0.62	5057
0	0.30	0.00	0.01	1435
1	0.81	0.92	0.86	13835
accuracy			0.77	20327
macro avg	0.59	0.50	0.50	20327
weighted avg	0.73	0.77	0.74	20327

From the above table, we can say that

- Correct Detection of Negative sentiment (-1) is 0 out of 5057 data
- Correct Detection of Positive sentiment (1) is 0 out of 13835 data
- Correct Detection of Neutral sentiment (0) is 0 out of 1435 data

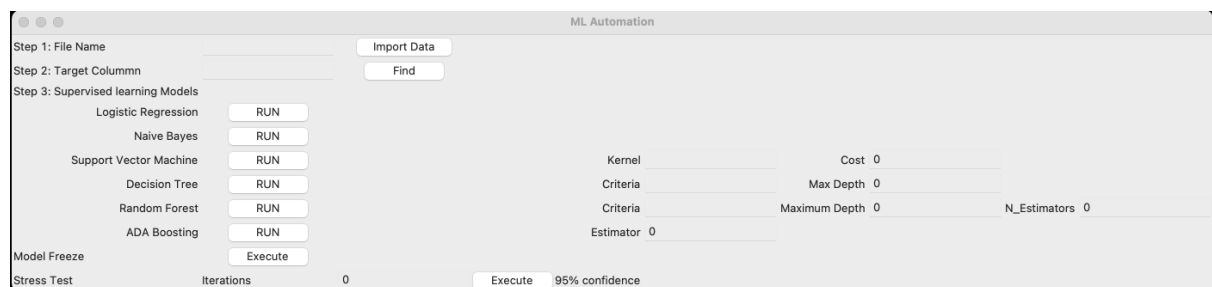
Furthermore, the ADA Boosting model gives us the accuracy score of 77%. Though the algorithm is so powerful, It helped in finding the best prediction for the feature of only positive sentiment.

Additionally, the only way to develop the model is to add some new data and see whether the negative and neutral sentiment have higher in numbers. Only when the data is not biased to one category, the model can be developed in terms of performance metric.

10.17 AUTOMATING THE MODEL CREATION:

To automate ML activities and deploy AutoML solutions, an ML pipeline is utilised. The pipeline begins with the collection of raw data, which is subsequently built to meet the needs of the algorithm and domain by utilising data cleaning and feature engineering techniques.

However, I have automated only the model creation part because in this generation we cannot depend on just one model. Hence, to play with hyperparameters and check the accuracy of different models, an automated GUI is used. In this process, we do not have to type the code, again and again, to check which model gives the best accuracy. The automated GUI would have the set buttons and a tabular column in which we have to ingest the data, and the automation would itself give the accuracy for each model.



The Automation would run as follows

Step 1: The first step is to put the file name of the data in the blank of the 1st row empty blank and then click the import data.

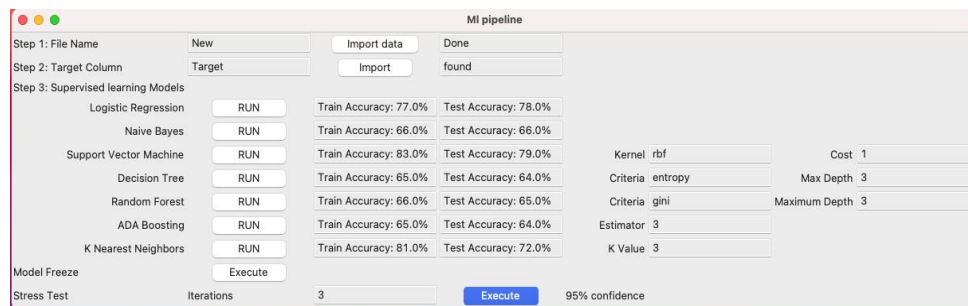
Step 2 : The second step is to write the target column name in the target column blank. If the target column is present inside the uploaded data, then it will run. Otherwise, it will throw an error. Additionally, if the target column is presented, then the data will be split into train and test categories with the amount of 20% in test data.

Step 3: By clicking the "run" button near each model, the model would give us the accuracy percentage, from which we can conclude which model gives us the best accuracy. Additionally, we can change the hyperparameters of each model and develop the model in terms of reducing the overfit and underfit issues.

Step 4: The model freeze would help us to save our model as a pickle file so that it can be sent to a machine learning engineer and he can deploy it to any of the cloud services so that they can run the model again whenever they get future data.

Step 5: The stress test confirms in showing that even if you give 100 models for this particular data, with a confidence of 95%, the accuracy would be between a particular number and a particular number.

Solution:

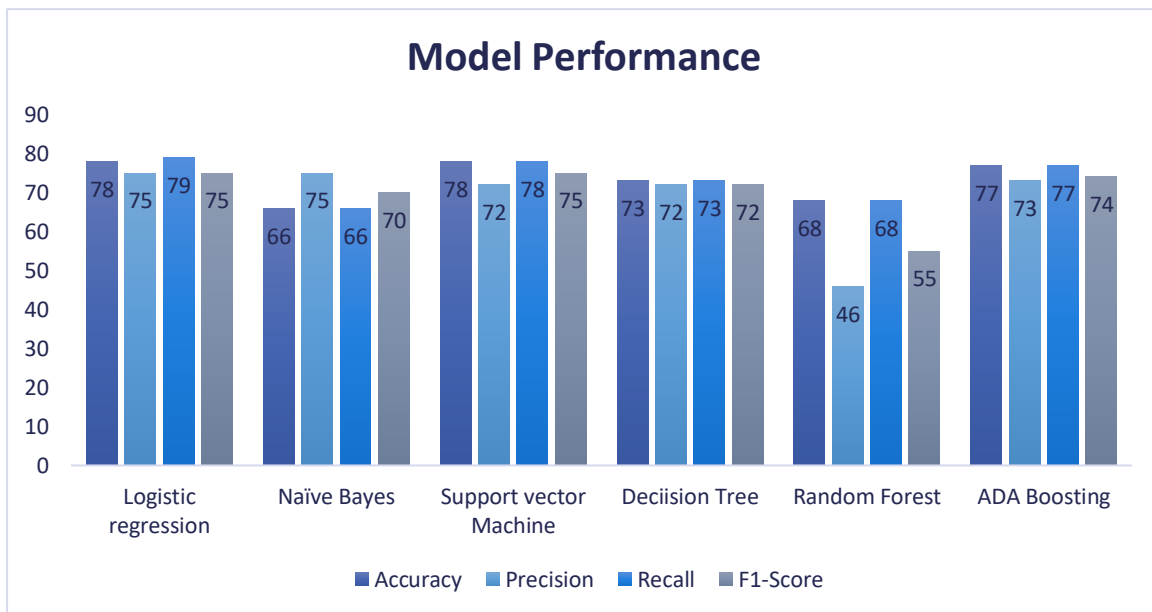


If we see the above GUI, the automation results are similar to how we created the models manually. By seeing the accuracy of the automation model, Logistic regression and SVM gave the better performance than compared to other models. Additionally, I tried the KNN model in automation, which gave 81% training accuracy. However, the testing accuracy is only 72%, which means it has over-fitting issues. Other models have an accuracy of between 65% and 66%, which is not up to the mark for production. Additionally, as I said above, we can save the model as a pickle file by clicking the model freeze button and deploy the machine learning model for further examination.

11.RESULT AND OUTCOME:

The classifiers and suggested model were used to detect the star rating of an Amazon mobile phone product, and the performance of the classifier was attained. In this section, classification measures will be compared to assess intervention.

Models	Accuracy	Precision	Recall	F1-Score
Logistic regression	78	75	79	75
Naïve Bayes	66	75	66	70
Support vector Machine	78	72	78	75
Deciision Tree	73	72	73	72
Random Forest	68	46	68	55
ADA Boosting	77	73	77	74



According to the comparison bar chart, the performance of the suggested model is the best among the selected classifiers. The Logistic regression model, in particular, outperforms the state-of-the-art ensemble model ADA Boosting. As a result, the performance of the Logistic regression, SVM is highest and best when compared to the selected classifiers and models in machine learning.

12. CHOOSING THE MOST EFFECTIVE MODEL

According to the experimental results, the accuracy of the Logistic regression, SVM, and ADA boosting classifiers is 78%, which is comparable for the three models. However, by examining the precision, recall, and F1 score, we can determine that the Support vector machine works quite well for this specific data. Furthermore, we may deduce that in the future, if we have additional data, we can

develop the SVM model and ensure that it has an accuracy close to 95%. As a result, we can select SVM model to deploy the model.

13. CONCLUSION:

13.1 RESEARCH PROJECT OBJECTIVES

With the conclusion of this study, all of the objectives were met, and the project will add to the current literature on the use of sentiment analysis to forecast star rating.

13.2 PROJECT PROBLEM STATEMENT

This study project was able to examine the new digital age's contribution to the subject of reputation measuring in market research. The study project supplied information on the feasibility of adopting new novel data collection and analysis tools as alternatives to established research methods such as surveys, in-depth interviews, focus groups, and so on. The strategies and approaches used in the study project addressed all of the issues encountered in market research, such as using automation for machine learning models.

13.3 EVALUATION:

The Amazon dataset contains far more information than similar rating-based datasets, such as the Netflix dataset. It enables us to investigate the relationship between each characteristic and the rating and use the relationships to forecast more correctly. We trained numerous models in this study, including logistic regression, random forest, and Decision tree models.

We also used text mining techniques like sentiment analysis to generate our features. To tailor the performance and regulate the model complexity, we compare the parameter settings for each model. Finally, we compare the models' performance on the test dataset.

The results reveal that the random forest model beats the others since it incorporates certain plausible characteristics drawn from the dataset's rich information. Furthermore, the automation performed well in terms of efficiently running the model. It also aids in giving the answer in a computer-efficient manner, allowing us to work with larger amounts of data in the future.

14. FUTURE WORK

Due to time constraints, we were unable to test various models such as neural networks and Stochastic Gradient Descent. In the future, we will also try to create more text mining models, such as bigram and trigram, and we may forecast the review star only based on the review text.

Furthermore, Amazon may provide further datasets that cover more firms from other market regions. Then we may examine user sentiment in various countries. Simultaneously, we could apply our review star prediction to all businesses, not only mobile phone brands.

Moreover, we may strive to reduce the product's negative sentiment by making suggestions to the seller of the mobile phone brand owner in amazon. For example, as I was carefully reviewing the unfavourable comments, I saw comments like

- Mobile is broken,
- USB cable not working
- Specification wrong on the amazon website

All of the preceding criticisms have been divided into three categories: damage, defective, and incorrect specification.

- Regarding the **damage** issue, I recommend that the vendor ship the goods with double bubble wrap rather than one, so that the product does not break during delivery.
- In the case of a **defective** product, I recommend that the vendor verify the goods twice or three times before shipping it to Amazon.
- In the case of a specification issue, the vendor can send the correct product specification to an Amazon employee so that the catalogue can be updated with the correct product specification.

If we offer this task to the seller, the negative sentiment will undoubtedly decrease, and the brand owners will not be churned out of amazon.

In terms of future automation work, we can attempt to automate the entire process in the future. For example, from data collection to model deployment. This allows us to save time when developing the model and will be useful in the future when running different sets of data.

15. REFERENCES

1. From Wikipedia, the free encyclopedia, Amazon(Company), Published in June 2022
2. Addie thomas, Mobile Phone and Smart Phone Market – Global Industry Analysis and Forecast 2015 – 2021, Published on July 3, 2015
3. Scott D. Anthony, Why Would Amazon Want to Sell a Mobile Phone?, Published on June 16, 2014
4. Monkey learn, Sentiment Analysis: A definite Guide, published in 2008
5. Vishal Shirsath, Sentiment Analysis of Events from Twitter Using Open Source Tool, Published in may 2016
6. H2O.ai WIKI, What is target variable in machine learning, published in 2001,2002

7. From Wikipedia, the free encyclopedia, Word2vec, Published in 2013.
8. Jacob Eisenstein, Unsupervised Learning for Lexicon-Based Classification, Published in November 2016.
9. Mohammad Darwich, Corpus-Based Techniques for Sentiment Lexicon, Published in October 2019
10. Python, what is python, published in 2001,2002
11. Andrew Luashchuk, Why python is perfect for machine learning and artificial Intelligence, Published on may 29, 2019
12. Michael F , Alteryx Alumni (Retired), Predictive process step, Published on February 5, 2019
13. Wael Etaiwi* and Ghazi Naymat, The Impact of applying Different Preprocessing Steps on Review Spam Detection, (EUSPN 2017)
14. Luthfi Ramadhan, A short introduction to TF-IDF vectorizer, Published on January 20,2021
15. Simplilearn, An introduction to logistic regression in python, Published on August 29, 2022
16. Dr. Jon Starkweather and Dr. Amanda Kay Moske , Multinomial logistic regression, 2022
17. Aman Kharwal, Classification report in machine learning, published on July 7,2021
18. Berrar, Daniel. 2018. "Bayes' Theorem and Naive Bayes Classifier." Encyclopedia of Bioinformatics and Computational Biology, vol. 1, Elsevier, pp. 403-412. Accessed 2022-02-07.
19. Wu, Lin and Weng, "Probability estimates for multi-class classification by pairwise coupling", JMLR 5:975-1005, 2004.
20. Geeksforgeeks, Decision tree Classifier, published on August 22, 2021, Improved By : bshyamanth, om_agarwal_2411, jaintarun, anyonepigwx, hardikkoriintern
21. Random Forest algorithm classifier
22. Anshul Saini, AdaBoost Algorithm – A Complete Guide for Beginners, Published On September 15, 2021
23. Mengqi Yu UC San Diego A53077101, "Restaurants Review Star Prediction for Yelp Dataset",2022