



اُونِيْوَرْسِيْتِيْ تِيْكْنُوْلُوْجِيْ مَآرَا
UNIVERSITI
TEKNOLOGI
MARA

FACULTY OF COMPUTER AND MATHEMATICAL SCIENCE

**ITS480
BUSINESS DATA ANALYTICS**

PROJECT TITLE:

REVIEW ON IPHONE 12

GROUP:
D2CS2416B

PREPARED BY:

NO	NAME	MATRIC NO
1.	MUHAMMAD ISMAIL BIN ELIAS	2019338721
2.	NASIHA BINTI ZAILAN	2019542219
3.	NUR AIN SYUHADA BINTI KAMARUZALI	2019542433
4.	SITI NUR FATIAH BINTI MOHAMAD NASIR	2019326663

PREPARED FOR:
MRS. MARHAINIS BINTI JAMALUDIN

SUBMISSION DATE:
6 February 2021

TEAM PROFILE



MUHAMMAD ISMAIL BIN ELIAS
NICKNAME: MAIL



**NUR AIN SYUHADA BINTI
KAMARUZALI**
NICKNAME: AIN



NASIHA BINTI ZAILAN
NICKNAME: SIHA



**SITI NUR FATIAH BINTI
MOHAMAD NASIR**
NICKNAME: TIHAH

TABLE OF CONTENTS

CONTENTS	PAGE NUMBER
1.0 INTRODUCTION	
1.1 Background of the Study	1
1.2 Research Statement	3
1.3 Significance of the study	3
2.0 BUSINESS UNDERSTANDING	
2.1 Determine business objective	4
2.2 Assess the situation	4
2.3 Determine the data mining goal	5
2.4 Performance Measure	5
3.0 DATA UNDERSTANDING	
3.1 Data sources	9
3.2 Data scrapping	10
4.0 DATA PREPARATION	
4.1 Manual data cleaning	14
4.2 Pre-processing – Data preparation and cleaning in Rapid Miner	15
5.0 MODELLING	
5.1 Model	23
5.2 Cross validation	26
5.3 Modelling process in rapid miner	27
5.4 Model comparison	28
6.0 FINDINGS	
6.1 Preference of user to purchase iPhone12	29
6.2 Rating of iPhone12 given by the user	30
6.3 Feedback from the user regarding iPhone12	31
7.0 CONCLUSION	32
REFERENCES	33
APPENDIX	34

1.0 INTRODUCTION

1.1 Background of the study

A smartphone is a mobile device essentially for communication that allows human interaction for the user to make and receive a call and to send text messages. As technology evolves, mobile growth has advanced to provide consumers with more features that can ease their job and daily task that is aligned with the modern lifestyle. Nowadays, many individuals use a smartphone to engage with friends, family and brands on social media. There are a variety of smartphone device brands that are popular on the market, such as Apple, Samsung, and Huawei (Global Smartphone Market Share: By Quarter, 2020). Following the upward trend of the Asian-Pacific smartphone industry, the demand for Malaysia's smartphone market is projected to remain steady at around five per cent annually in the coming years. Approximately 17.2 million smartphone users were present in the country in 2018. This number is expected to rise to almost 21 million by 2023 (Statista, 2020).

According to the Global Smartphone Market Share: By Quarter (2020), Apple company has dominated the smartphone market industry at the end of 2019, making up for 18% due to the launch of iPhone 11 series across all regions. While other mobile devices come up with operation system of Android, Apple stands out for its exclusive operating system, IOS. Interestingly, the iPhone sales performance is always high despite the price gap, hardware and features as well as its customization with the Android's smartphone. Iphone is famous for their privacy and security concern, together with the simplicity of use that makes it convenient among their users. According to the Reinfelder et. al (2014), iOS malware is rare compared to Android because all apps in the App Store undergo a review process in order to ensure that the apps work according to their description. This also means that the apps should not have malicious functionality. Furthermore, IOS system radically changed the handling of personal data where the users now have to give runtime consent for many more data types, such as contacts, calendar, photos, Twitter or Facebook accounts. (Reinfelder et. al, 2014).

Nowadays, iPhone smartphone is popular among Malaysian users and becoming such a trend where secondhand iPhone can be found widely marketed across ecommerce sites and local shops. Since the first arrival of Apple's iPhone on June 29, 2007, it has sold astoundingly for four million units (Carew, 2008). Apple is a well-known company for its great marketing strategy. Furthermore, iPhone only release a set of new model series once a year make it as the top brand competing with the other smartphone companies. In order to attract new buyers and current users to upgrade, Apple has to come up with new phone series every year, catering to all sorts of budgets and users. Recently, the world is facing severe economic issue. Thus, this also impacts the financial situation in Malaysia. However, Apple has been launching four new iPhone 12 model series which are iPhone 12, iPhone 12 mini, iPhone 12 pro and iPhone 12 Pro Max in September 2020 and the premises that selling Apple products were raided by the public. Hence, the craze of the public to get those iPhone 12 series remains unchanged. The latest features of iphone12 have brought the attentions from the audiences with the introduction of 5G technology, A14 Bionic processor and also come up with better camera. The price for iPhone 12 mini starting from MYR3399, MYR3899 for iPhone12 model, MYR4899 for iPhone 12 pro model and MYR5299 for iPhone 12 pro Max model, according to the Apple authorized reseller website, the Switch. Therefore, this study intends to know about the user feedback of iPhone 12. The data will be obtained from the user reviews via multiple online websites.

1.2 Research Statement

Apple's product considered by many to be a high-class brand. Hence, with the new launch of iPhone12, how does it affect the image of Apple, does it retain, increase the company reputation, or make it worse? That is the question that piques our interest. To answer that question, we will study the rating and opinion of iPhone12 buyers. This way we can identify how the users feel about the newer model. The relationship between rating and the image of Apple is quite intriguing because the rating and user review may affect the potential selling of future product as people may refuse to buy the newer one.

Another reason to conduct this study is due to its sales performance. Despite launching new model once, a year and selling it at high price, Apple has generated high profits. This demonstrates how strong the Apple brand is. To be specific, this strategy is quite opposite of what Apple's competitors are doing. Other company like Vivo, Oppo and Xiaomi sometimes release several models in 1 year with much affordable price. The reason Apple is doing this kind of thing is to ensure the exclusivity and the quality of the smartphone manufactured. Because of its exclusivity, people equate the use of an iPhone with a particular social class.

1.3 Significance of the study

Analyzing Apple's latest product could benefit the potential buyers particularly. This is because there may be among us who are interested in buying a new phone but do not know which model to purchase. Since the iPhone12 has just been launched, we can know if it is worth spending our money on the new Apple product.

2.0 BUSINESS UNDERSTANDING

Business understanding refers to the first stage in Cross-industry standard process for data mining (crisp-dm). This is the fundamental phase before data mining techniques can be conducted. Throughout this step, it helps the researcher to Understand the project objectives and requirements from a business perspective, and then convert this knowledge into a data mining problem definition so, that a preliminary plan can be designed to achieve the objectives (Data Mining Process, 2021). This phase involves 4 tasks that must be followed which are determine business objective, assess situation, determine data mining goals and measure the performance.

2.1 *Determine business objective*

This process refers to the description of primary objective that researcher want to achieve from a business perspective. Henceforth, this study is intended to achieve the business objectives that follows:

1. To analyze the preference of likelihood to purchase iPhone 12.
2. To analyze the review or feedback from the users.
3. To determine the rating of iPhone 12 given by the users.

2.2 *Assess the situation*

Before sentiment analysis can be conducted, it is required to study and investigate about the resources needed in analysis, the constraint that may restrict the study and other factors that need to be considered in determining data mining goals and project plan. Therefore, throughout this study, the data mining technique is performed by conducting sentiment analysis to detect the polarity of emotions and opinion of the user reviews about iPhone12. Appropriate data mining technique should be applied as wrong approach may lead to inaccurate result for prediction. Other method, steps and software that will be used in this study should be taken with care because this study utilize machine learning technique for modelling. With the use of sentiment analysis is applied in this study, there are only certain tools that are available for text-mining processing to collect the data, while for data cleaning, relevant software like the comma-separated value (CSV) and the Microsoft excel software is suitable as the use of this tools is widely known for its functionality that is easy to use by many data analysts for data management process like storing and cleaning the data. Finally, for data analysis process, there are limited relevant software and package that available to analyze a large volume of data that suitable according to its datatype.

2.3 *Determine the data mining goal*

This step is aimed to translate the business question to data mining goals. To be specific, this stage intended to achieve the expected outcome in technical terms. For example, in this study, there are 3 data mining goals that the researcher aimed to achieve that follow:

1. To determine the preference of likelihood for sentiment polarity detection of iPhone12.
2. To identify the most frequent word used among the user reviews of iPhone12.
3. To identify the highest frequency of rating given by the iPhone12 reviewers.

2.4 *Performance measure*

This final stage is important where the proper planning and methodology of analysis is necessary to gain the expected outcome, which in this study, to detect the polarity of opinion among user reviews of iPhone12. Data mining technique and modelling should be taken with care to find the best model either Naïve Bayes, support vector machine or random forest.

Sentiment analysis

Sentiment analysis refers as opinion mining. It is a natural language processing (NLP) technique that recognizes the emotional tone behind the body of the text. This is a common way for organizations to recognize and categorize product, service, or concept opinions. This involves the use of data mining, machine learning (ML), artificial intelligence (AI) to mine text for sentiment and subjective knowledge (TechTarget, 2019).

Sentiment analysis systems help the organization gain information from the unstructured text from online sources such as emails, blog posts, ticket assistance, webchats, social media networks, forums, and comments.

Sentiment analysis is the method of identifying positive or negative emotions in the text. Businesses are also used to detect emotions in social data, to gauge brand credibility, and to understand consumers (Everything There Is to Know about Sentiment Analysis, 2020).

Sentiment analysis is extremely useful because it helps companies to easily understand the overall opinions of their customers. It can make faster and more accurate decisions by automatically sorting the emotions behind scores, social media conversations, and many more.

Text mining pre-processing technique

Data mining is used to find useful information on a large amount of data. Data mining methods are used to apply and solve various forms of research problems. Data mining related research areas include text mining, web mining, image mining, sequential pattern mining, space mining, medical mining, multimedia mining, structure mining, and graphic mining. Text mining is the method of extracting valuable information from text records. It is also referred to as knowledge discovery in text (KDT) or knowledge of intelligent text analysis.

Text mining is a technique that gathers information as well as pattern-finding from both structured and unstructured data. In a variety of research fields, such as natural language processing, information retrieval, text classification and text clustering, text mining techniques are used. The steps in the text mining pre-processing technique are shown in the Figure 2.1.

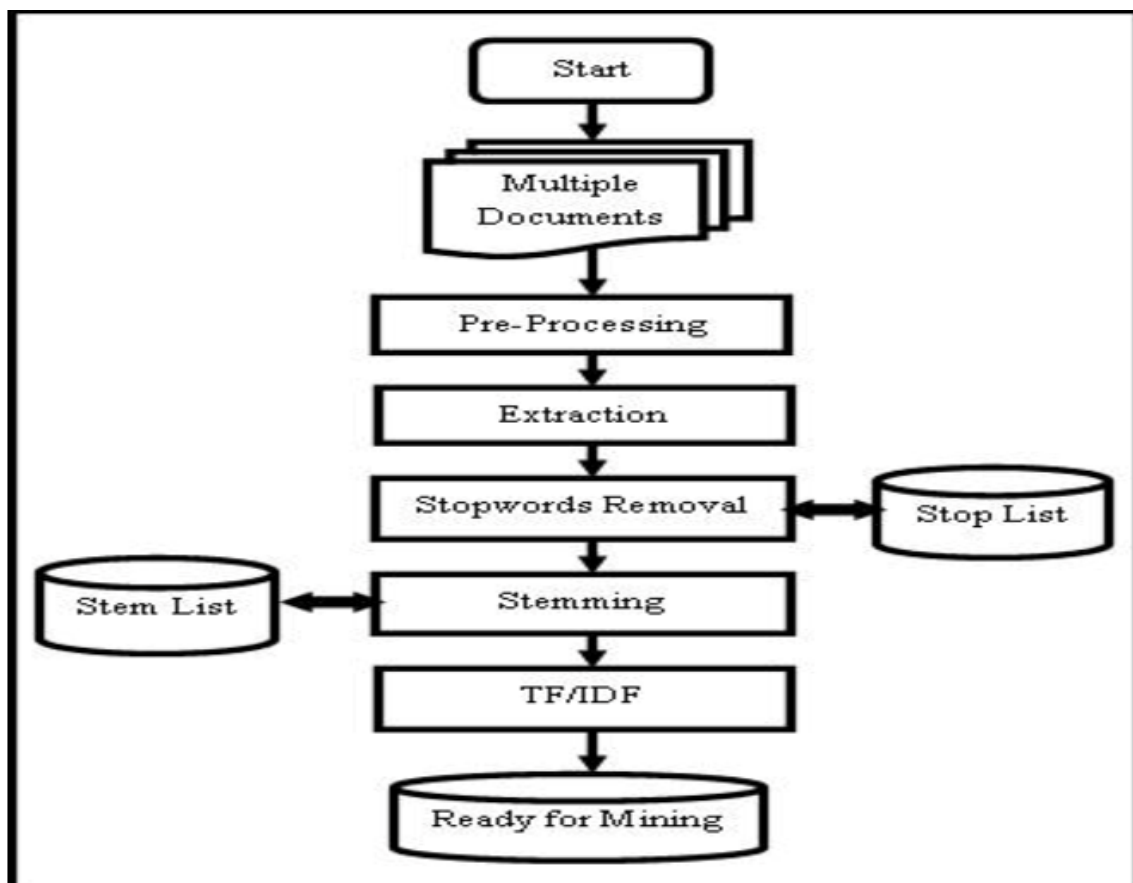


Figure of Text mining pre-processing techniques

The process in Sentiment Analysis

Step 1: Monitor iPhone12 websites review.

- To address the sentence in the speech components and clarify the syntactical relationship, the sentence must be parsed in order to obtain the actual content. The aim of this parsing is to generally enforce the structure on semi-structured data.
- The structure must be adequate to define the portion of the raw text of the actual content of the review, the title, the date of the review, and others. The output is a series of phrases and terms that speak of the product of interest.

Step 2: Collect the reviews.

- The raw data involved have been extracted and translated into an appropriate representation of the text. This is intended to reflect the array of phrases and terms in a standardized manner for downstream analysis and to measure the number of times a term occurs.
- Corpus is a collection of documents. In order to be able to conduct a search for future reference and research, the corpus was represented to archive them.
- Reverse indexing provides a way for all documents containing a specific feature to keep track of the list.
- Documents are often relevant only in the context of a corpus or a specific collection of documents. Therefore, classifiers need to be trained on a specific set of documents. Any changes to the corpus will require the retraining of the classifier.
- Corpus changes continuously over time and not only adds new documents, but the distribution of words can change over time. This could reduce the efficiency of classifiers and filters if they are not retained.

Step 3: Sort the reviews

- Once all the reviews have been collected and represented, sort them by subject matter of interest (feedback of iPhone12) which focuses on attribute name, the rating and the comment. The reviews must be classified in order to sort them out.
- Typically, the subject tag often involves having a team of human users determine the classification of the review and tag it accordingly.

Step 4: Determining type of review (good or bad).

- Sentiment analysis is conducted by analyzing the reviews of iPhone12 from multiple websites. From the review, the insight about the opinion of using iPhone12 is obtained.
- Search by relevance is often made possible using Term Frequency- Inverse Document Frequency (TF-IDF). The frequency – inverse document frequency is a weight- based metrics to identity reviews/ documents relevant to some query terms.
 - I. Document frequency: Provides information about how many documents that contains a term.
 - II. IDF: Provides information about how rare the term is across the document corpus and measure the relevance that DF does not provide.

Step 5: Assess model classification for best accuracy

- To determine if a review is good (positive) or bad (negative), classifiers include Naïve Bayes, Support Vector Machine (SVM) and Random Forest were used.
- Naïve Bayes: A family of probabilistic algorithms that uses Bayes' Theorem to predict the category of a text.
- Support Vector Machines: A non-probabilistic model which uses a representation of text example as points in a multidimensional space.
- Random Forest: Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction.
- A major bottleneck of this step is the need for tagged training data. Three approaches for this step are to identify good and bad review and utilize sentiment dictionary.

In conclusion, by automating business processes, generating actionable insights and saving hours of manual data processing, the sentimental analysis system helps companies make sense of unstructured text. In other words, by making the terms more effective.

3.0 DATA UNDERSTANDING

The second stage after business understanding for the Cross-industry standard process for data mining (crisp-dm) is data understanding. This process starts with the initial data collection and continues with activities in order to get acquainted with the data, to identify data quality issues, to discover first insights into the data or to detect interesting subsets to develop hypotheses for hidden information.

3.1 Data sources

In this study, the dataset about iPhone12 reviews were used to detect the polarity of opinion among the iPhone12 reviewers. The data is obtained by scrapping the observations from multiple websites review online. There are 5 variables that was obtained throughout the scrapping process which the raw dataset included variables like the web-scrapper-order, web-scrapper-start-url, Name, Rating and the comment. However, for sentiment analysis, only 3 variables will be used which are the name, the rating and the comment. There are 1659 observations that were obtained from scrapping multiple website reviews.

AutoSave

amazonreview-1 - Read Only

Search

nashuhalan

File Home Insert Page Layout Formulas Data Review View Help Power Pivot

Cut Copy Paste

Format Painter

Clipboard

Calibri 11

B I U

Font

Wrap Text

Align Center

Alignment

General

%

Number

Conditional Formatting

Format as Table

Styles

Cell Styles

Insert

Delete

Format

Cells

AutoSum

Fill

Clear

Editing

Sort & Filter

Find & Select

Analysis

Data Analysis

Sensitivity

Sensitivity

Share Comments

H3

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	web-scrap	web-scrap	name	rating	comment																	
2	16124039f	https://ww	Keshav	5.0 out of	Well i dont have to say much because my phone is working perfect so i would just say go for it if you really think you can buy this phone. Dont worry its worth having it. Its been almost 1 month and i havent face a																	
3	16124039f	https://ww	Naku	4.0 out of	Very op																	
4	16124039f	https://ww	Sarita shar	5.0 out of	Waaah																	
5	16124039f	https://ww	Abhinay d	5.0 out of	Its osun phone.i																	
6	16124039f	https://ww	G narayan	5.0 out of	Camera quality is excellent, battery life is also ok. Compare to android not much. But good recently quite. I am first time user, apple company sells now separate every part business like . This is not corect																	
7	16124039f	https://ww	Jaswinder	5.0 out of	best phone																	
8	16124039f	https://ww	sanam s.	5.0 out of	Excellent product																	
9	16124039f	https://ww	Ananthakr	5.0 out of	Great device. Amazing Cameras. Battery is good for a day. I don't play games so i'm not sure about it.																	
10	16124039f	https://ww	Amitesh	5.0 out of	Good but not good as iPhone 12 Pro nice flagship nice phone feels like holding iPhone 4+ iPhone 11 mixed																	
11	16124039f	https://ww	Mukund	5.0 out of	Okkk so done with one kidney going to sell another one for power adapter and for case and ofcourse value for kidney																	
12	16124039f	https://ww	Ashik K.	5.0 out of	Excellent device, beautiful IOS, AMOLED display and design. The only drawback for me coming from Android devices with massive batteries is that this only has a single day battery while i am used to more.																	

Figure of example dataset that has been scrapped in the Webscrapper.io

The variables for the scrapped dataset in Webscraper.io and its data type:

Attributes	Data type
Web-scrapper-order	Numerical
Web-scrapper-start-url	Nominal
Name	Nominal
Rating	Numerical
comment	Nominal

3.2 Data scrapping

For data collection method, web scrapping is applied by using Webscrapper.io package from the internet. web scraping refers to the extraction of data from a website. The data is extracted and will be used for sentiment analysis in this study. Webscraper.io is a free Google Chrome web browser extension that enables users to use HTML and CSS to extract information from any public website. In this study, Webscraper.io extract user reviews of iPhone 12 from multiple website sources. There are 11 review websites that were used for collecting data. Hence, in this study, there are several steps involved for data scrapping of iPhone12 user reviews.

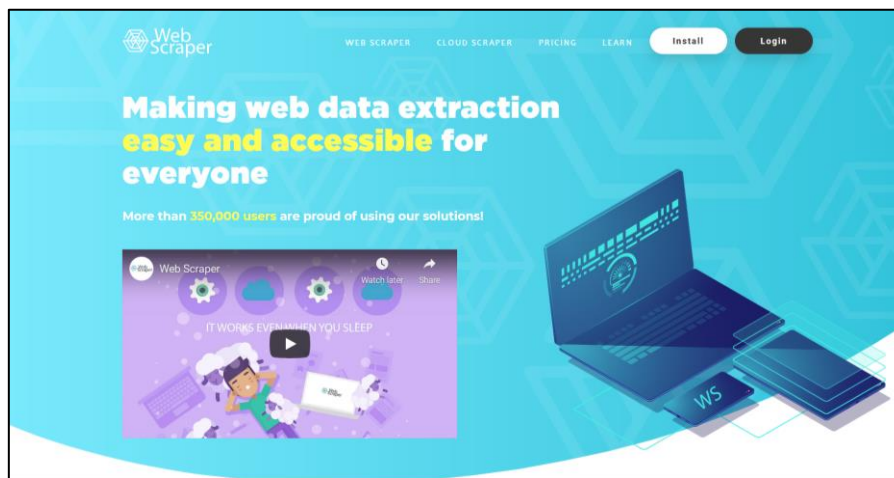
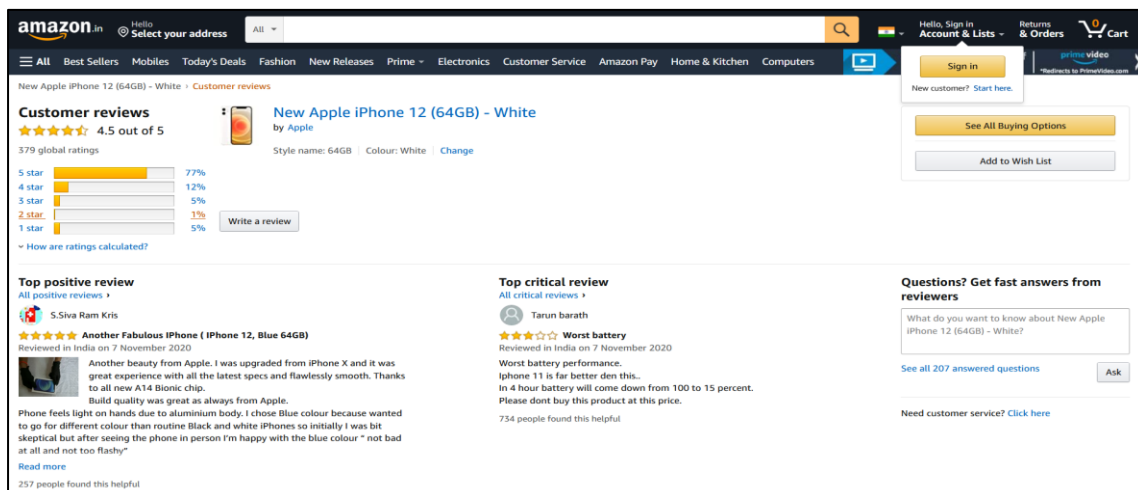


Figure of webscrapper.io software

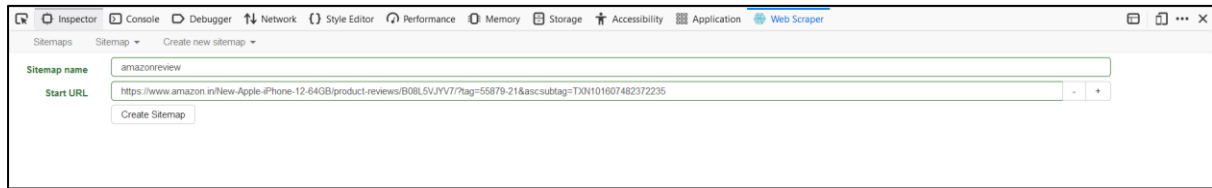
Data scrapping process:

Step 1: Open the website link where data scraping can be performed in this stage. For example, in this study, amazon website is used for extracting the reviews of iPhone12.

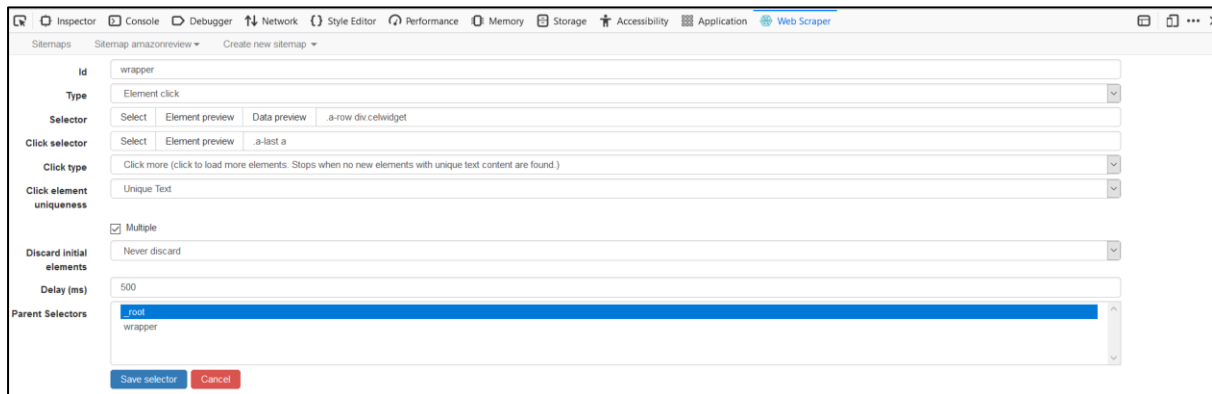
The link: <https://www.amazon.in/New-Apple-iPhone-12-64GB/product-reviews/B08L5VJYV7/?tag=55879-21&ascsubtag=TXN101607482372235>



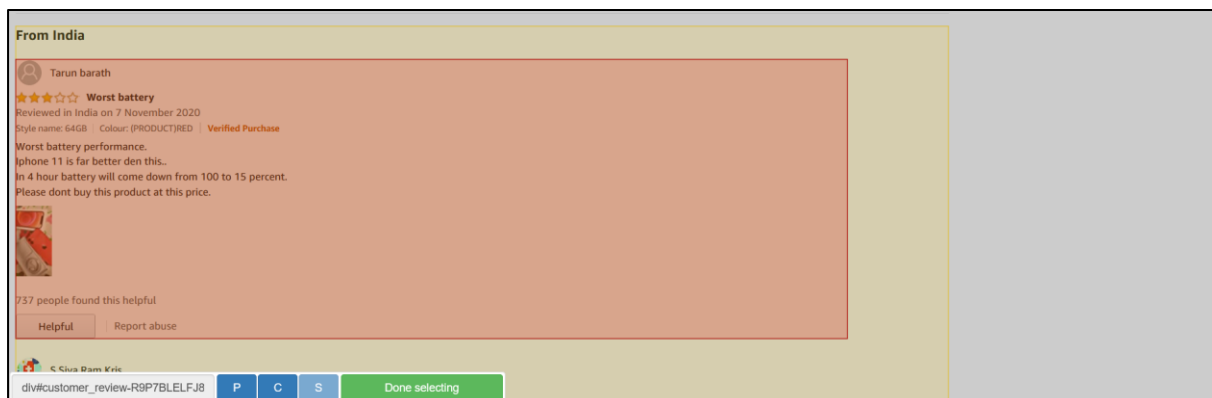
Step 2: Creating new sitemap




Step 3: Create the wrapper by selecting parent element where click type used for this process is 'click more'.



This includes the selection of data such as the name, the rating and the comment in one column.



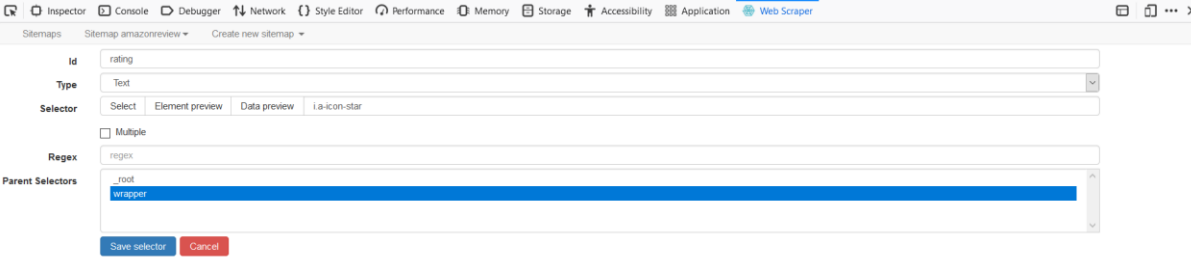
Step 4: Create the child element selector for id name.



Id: name
Type: Text
Selector: Select Element preview Data preview span a-profile-name
Multiple: ☐
Regex:
Parent Selectors: _root, wrapper
Save selector Cancel

Tarun barath
★★★★☆ **Worst battery**
Reviewed in India on 7 November 2020
Style name: 64GB | Colour: (PRODUCT)RED | **Verified Purchase**
Worst battery performance.
Iphone 11 is far better den this..
In 4 hour battery will come down from 100 to 15 percent.
Please dont buy this product at this price.

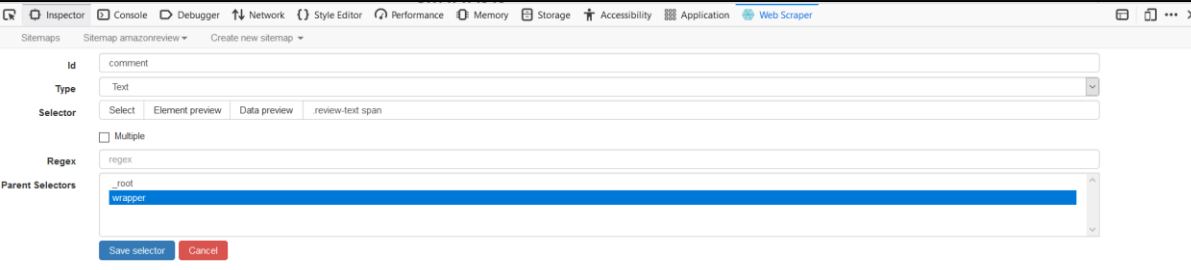
Step 5: Create the child element selector for id rating.



Id: rating
Type: Text
Selector: Select Element preview Data preview i a-icon-star
Multiple: ☐
Regex:
Parent Selectors: _root, wrapper
Save selector Cancel

Tarun barath
★★★★☆ **Worst battery**
Reviewed in India on 7 November 2020
Style name: 64GB | Colour: (PRODUCT)RED | **Verified Purchase**
Worst battery performance.
Iphone 11 is far better den this..
In 4 hour battery will come down from 100 to 15 percent.
Please dont buy this product at this price.

Step 6: Create the child element selector for id comment.



Id: comment
Type: Text
Selector: Select Element preview Data preview review-text span
Multiple: ☐
Regex:
Parent Selectors: _root, wrapper
Save selector Cancel

Tarun barath
★★★★☆ **Worst battery**
Reviewed in India on 7 November 2020
Style name: 64GB | Colour: (PRODUCT)RED | **Verified Purchase**
Worst battery performance.
Iphone 11 is far better den this..
In 4 hour battery will come down from 100 to 15 percent.
Please dont buy this product at this price.

step 7: Scrape the data by selecting scrape button

The screenshot shows the 'Selectors' panel in the Web Scraper application. The panel lists three selectors: 'an a-profile-name', 'icon-star', and 'view-text span'. Each selector has a 'type' of 'SelectorText', 'Multiple' set to 'no', and a 'Parent selectors' of 'wrapper'. The 'Actions' column for each selector includes buttons for 'Element preview', 'Data preview', 'Edit', and 'Delete'.

ID	Name	Selector	Type	Multiple	Parent selectors	Actions
	an a-profile-name	an a-profile-name	SelectorText	no	wrapper	Element preview, Data preview, Edit, Delete
	icon-star	icon-star	SelectorText	no	wrapper	Element preview, Data preview, Edit, Delete
	view-text span	view-text span	SelectorText	no	wrapper	Element preview, Data preview, Edit, Delete

Step 8: Export the data in the format comma separated value (CSV)

The screenshot shows the 'Data' panel in the Web Scraper application. The panel displays a table of scraped data with columns: 'web-scraper-order', 'web-scraper-start-url', 'name', 'rating', 'comment', and 'title'. The table contains four rows of data.

web-scraper-order	web-scraper-start-url	name	rating	comment	title
1612403966-17	https://www.amazon.in/New-Apple-iPhone-12-64GB/product-reviews/B08L5VJYV7/?tag=556879-21&ascsubtag=9	shubham	5.0 out of 5 stars	Pros: there are 1000s of blog for pros Cons: Apple has to work to improve their battery life Charger 1 Charger 1 if apple really cares about environment so much then they must deliver a phone once in 3 years. Or, give the users a usb c. Don't rely on their screen advertising. Use a cover. Glass is still a glass and you can observe scratches after a month of usage. 5g isn't for India, still price in terms of dollars was same!	Apple ... has always been just apple.
1612403966-44	https://www.amazon.in/New-Apple-iPhone-12-64GB/product-reviews/B08L5VJYV7/?tag=556879-21&ascsubtag=9	Keshav	5.0 out of 5 stars	Well i dont have to say much because my phone is working perfect so i would just say go for it if you really think you can buy this phone. Dont worry its worth having it. Its been almost 1 month and i havent face any issue even battery backup is fine for me.	Best deal, Got it 128gb for 75900 🍌🍌
1612403966-138	https://www.amazon.in/New-Apple-iPhone-12-64GB/product-reviews/B08L5VJYV7/?tag=556879-21&ascsubtag=9	Naku	4.0 out of 5 stars	Very op	Op
1612403966-140	https://www.amazon.in/New-Apple-iPhone-12-64GB/product-reviews/B08L5VJYV7/?tag=556879-21&ascsubtag=9	Sarita sharma	5.0 out of 5 stars	Waaah	Amazing product

4.0 DATA PREPARATION

Data preparation is an important stage before sentiment analysis and modelling can be constructed. This stage consists of data cleaning where preparing the data required cleaning to make a good prediction on sentiment analysis. The data cleaning is conducted to remove any missing values and inconsistencies in the data.

4.1 Manual data cleaning

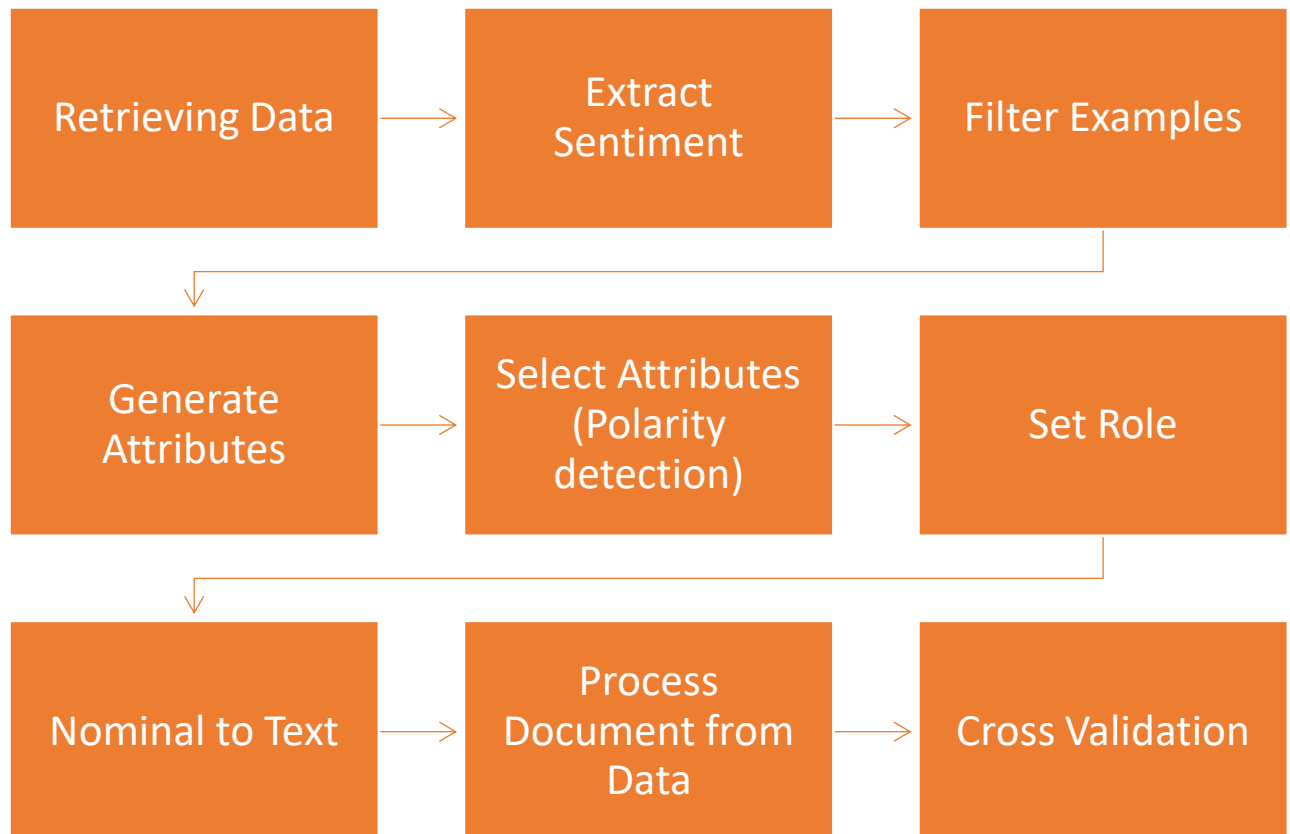
Before Polarity detection on the data of iPhone12 reviews can be performed, manual data preparation for training and testing dataset are done in the Microsoft excel worksheet to check for any inconsistencies and missing values, where missing values are identified as null in the observations. Firstly, all the observations that is obtained from scrapping are compiled together in one Microsoft excel worksheet format to prevent data loss and to conduct data cleaning. This study is heavily relied on the comment of reviews. Henceforth, data cleaning in Microsoft excel worksheet focuses more on cleaning attribute "comment". A row observation will be deleted if there are any missing values detected in attribute "Comment". After cleaning, there are 1161 of total observations left for analyzing the sentiment of iPhone12 user reviews. The clean data will be exported into the Rapid Miner software for further data cleaning and preparation of training dataset for sentiment analysis.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	web-scrap	web-scrap	Name	Rating	Comment													
2	16087442	https://wv	A Gupta	5	5.0 Rated Apple iPhone 12 as Excellent.													
3	16087442	https://wv	Alex Howa	1	Rated Apple iPhone 12 as Poor													
4	16087442	https://wv	amitabh sh	5	Rated Apple iPhone 12 as Excellent													
5	16087457	https://ind	Anuj Bhati	5	If you are looking to upgrade your existing phone and have a spare Rs 79,900 for a new smartphone, I won't stop you from buying the iPhone 12. You are													
6	16087442	https://wv	Arjun Prabh	null	They are charging a high price and not even providing a charging adapter & earphones. Better to buy older models !													
7	16087442	https://wv	Athem Sim	5	Rated Apple iPhone 12 as Excellent													
8	16087442	https://wv	Brendan Tr	5	the guy commenting above is a big fool. He should do his research before commenting. First of all this are the specs of the current iphone 11 series . This is													
9	16093845	https://wv	Abhi Kada	5	Amazing ♥													
10	16093845	https://wv	Abhishek t	1	Face id not proper working Bluetooth fully not working after 20days date of purchasing													
11	16093845	https://wv	Advait	5	All I phone models are premium phones, coming to 12 price is to costly, but the quality is awesome.Camera, display, processor and look of the body is pr													
12	16093845	https://wv	Ajay Malik	5	Awesome.. this is my first switch to an IOS Device after using an android device untill now. And i am happy. The built quality, display, performance and resp													
13	16093845	https://wv	Akhil Gauti	5	Amazing looks.													
14	16093845	https://wv	Alakshya	5	Perfect phone in every aspect. Look and feel amazing!!													

Example of compiled dataset in Microsoft Excel worksheet format

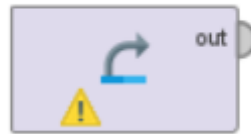
4.2 Pre-processing – Data preparation and cleaning in Rapid Miner

Data cleaning in Rapid Miner is the process of detecting and correcting or filtering corrupt or inaccurate observations from the document in the Rapid Miner. This is a crucial stage for preparing training dataset for the modelling. Therefore, data preparation and data cleaning steps in Rapid miner are as follows:

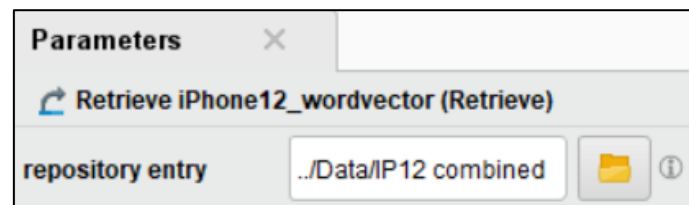


Stage 1: Retrieving Data

Retrieve IP12 Combi...



Retrieve operator is functioned to access stored information in the repository and load them into the process. The Retrieve Operator loads a RapidMiner Object into the Process. The parameter used for the operator refers to the path to the RapidMiner object which should be loaded. Method used in this study is “../Data/IP12 combined” that looks up an entry “IP12 combined” located in a folder “Data” next to the folder containing the current process. The data is imported from the Microsoft excel worksheet.



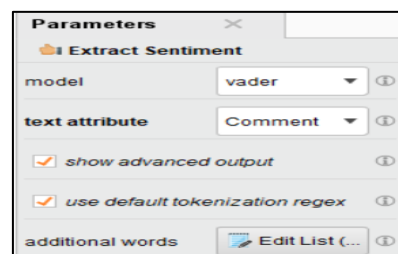
Open in Turbo Prep Auto Model					
Row No.	web-scrape...	web-scrape...	Name	Rating	Comment
1	1608744211-...	https://www.g...	A Gupta	5	5.0 Rated Ap...
2	1608744211-...	https://www.g...	Alex Howard	1	Rated Apple i...
3	1608744211-...	https://www.g...	amitabh shek...	5	Rated Apple i...
4	1608745711-...	https://indian...	Anuj Bhatia	5	If you are look...

Example of retrieved data in Rapid Miner

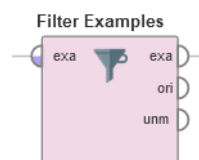
Stage 2: Extract sentiment



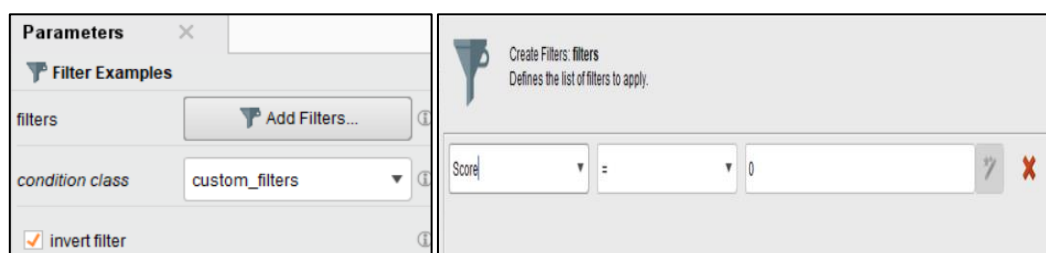
Extract Sentiment operator measures the sentiment based on what the dictionary contains in the Rapid Miner, which represents whether the sentiment is negative when the value is close to -1 or positive when the value is near to 1. The Vader model is chosen for the parameter in which Valer (Valence Conscious Dictionary for Sentiment Reasoning) is a model used for text sentiment analysis that is either positive or negative and sensitive to polarity. In this research, sentiment analysis is performed on the text attribute 'comment' to detect either "positive" or "negative" polarity.



Stage 3: Filter example



The filter example operator filters all the example input with the filter condition specified. This operator reduces the number of examples in example set but does not effect the number of attributes. The string matches all examples when the comment attribute contains the score of 0. The operator filters the data with such a regular expression, (Score = 0).



Stage 4: Generate attribute



The Generate Attributes operator constructs new attributes from the attributes of the input example set and arbitrary constants using mathematical expressions. New attribute is set with the attribute name is 'sentiment'. The sentiment value will display for "positive" and "negative". The function expression = $\text{if}(\text{score} < 0, \text{"Negative"}, \text{"Positive"})$.

 The configuration window for the 'Generate Attributes' operator is split into two panes. The left pane, titled 'Parameters', shows the operator name 'Generate Attributes', a button 'Edit List (1)...', and a checked checkbox 'keep all'. The right pane, titled 'Edit Parameter List: function descriptions', contains a table with two columns: 'attribute name' and 'function expressions'.

attribute name	function expressions
Sentiment	$\text{if}(\text{Score} < 0, \text{"Negative"}, \text{"Positive"})$

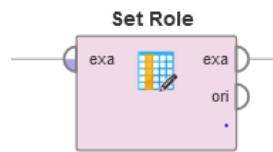
Stage 5: Select Attribute



This Operator filters the attribute by removing the non-selected attributes. By directly selecting the attributes to be taken for analysis, the selected attribute which are the "comment", the "name" and the "sentiment" are selected with the "Subset" attribute filter type is applied in the Rapid Miner process. Subset type is known as a combination of the select attributes operator and the subprocess operator.

 The configuration window for the 'Select Attributes' operator is split into two panes. The left pane, titled 'Parameters', shows the operator name 'Select Attributes', a dropdown menu for 'attribute filter type' set to 'subset', a button 'Select Attributes...', an unchecked checkbox 'invert selection', and a checked checkbox 'include special attributes'. The right pane, titled 'Selected Attributes', contains a search bar and a list of selected attributes: 'Comment', 'Name', and 'Sentiment'.

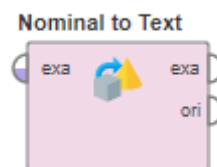
Stage 6: Set Role



This operator assigned special role for the attribute sentiment. With the aim of analysing the sentiment on iPhone12 reviews, the target role for the parameter is set to the 'Label' to assign the attribute 'sentiment' as the target variable or dependent variable for the polarity detection analysis.

Parameters	
Set Role	
attribute name	Sentiment
target role	label
set additional roles	Edit List (0)...

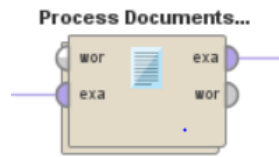
Stage 7: Nominal to Text



After setting the role for attribute 'sentiment'. The operator 'Nominal to Text' is used to convert nominal attributes to string attributes as this study is aimed to analyse the review of iPhone12 user. In Rapid Miner, if the data is in text format, it is considered as nominal data type and therefore, text mining cannot be performed. By using all attribute filter type, Rapid Miner selected all the attribute to convert its metadata in this study.

Parameters	
Nominal to Text	
attribute filter type	all
<input type="checkbox"/> invert selection	
<input type="checkbox"/> include special attributes	

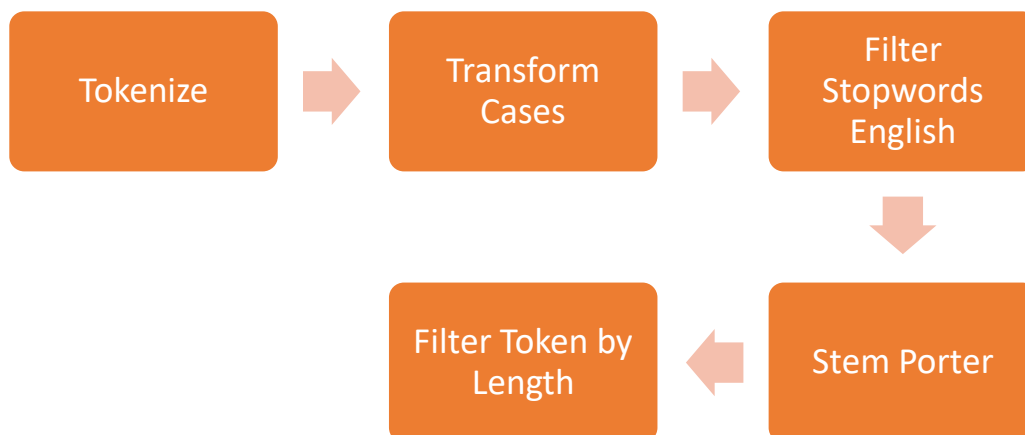
Stage 8: Process Document from Data



Process document from data operator outlines the steps necessary for text mining to complete the pre-processing of sentiment analysis in this study. The goal of this operator is to create word vector or a bag of words containing words from the attribute 'comment'. The use of TF-IDF for this operator helps to identify and evaluate relevant words in the document for a collection of documents.

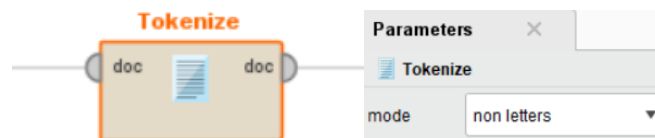
Parameters	
Process Documents from Data	
<input checked="" type="checkbox"/>	create word vector
vector creation	TF-IDF
<input checked="" type="checkbox"/>	add meta information
<input type="checkbox"/>	keep text
prune method	none
data management	auto
<input type="checkbox"/>	select attributes and weights

This operator has a process-in-process in the operator where there are 5 operators are added into the process which are tokenize, transform cases, filter stopwords (english), stem (porter) and filter token by length as figure below:



Tokenize (step 1)

Tokenize operators is functioned to split each of word characters into token, which will crate bag of words. Maximum range for tokenization is from 1 to infinity and it only filter non-letters text which means that the output will produce token for the characters only. In this step, any number values in attribute 'comment' will not be included for tokenization.



Transform cases (step 2)

Transform operator is used to standardize the token either uppercase or lowercase. This operator transform cases of characters in the document. In this step, parameter lower case is used to transform all the tokens into lower case form.



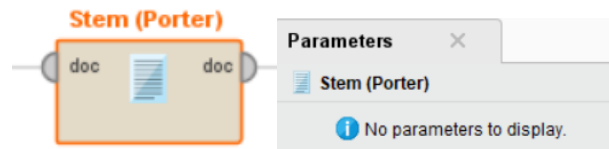
Filter stopwords English (Step 3)

In this step, filter stopwords (english) operator remove English stop words from a document by removing every token that equals to built-in stopwords list from the RapidMiner. Word like 'a',..... will be removed from the document. For this operator to work properly, every token should represent a single english word only.



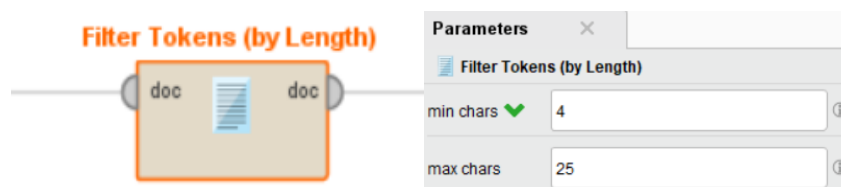
Stem Porter (Step 4)

In this step, this operator stems english word by removing affixes from the document that have similar term of word by stemming the token to the root english word. The goal is to reduce the length of words until minimum length is reached.



Filter tokens by length (Step 5)

In this step, this operator filter the tokens based on their length. Relevant length parameter is set to identify meaningful word from token starting from length 4 to maximum 25 characters.



5.0 MODELLING

5.1 Model

In this modeling phase, different supervised machine learning models are used. Overall, there are three models are deployed in this study namely as naïve bayes, support vector machine and random forest.

1. Naïve Bayes

Naïve Bayes algorithm follows the technique of classification which refers the “Bayes theorem” for classification. It assumes the independence within variables. According to the Dey et al. (2016) stated that Naïve Bayes was introduced under a different name into the text retrieval community and remains a popular baseline method for text categorizing and the problem of judging documents as belonging to one category or the other with word frequencies as the feature. An advantage of Naïve Bayes is that it only requires a small amount of training data to estimate the parameters necessary for classification.

Abstractly, Naïve Bayes is a conditional probability model. According to the Leung (2007) stated that Naive Bayes find the probabilities within class with the other classes. In simple terms, Naive Bayes assumes that particular aspect within class is not related to any other aspects within class, it is independent on its own. $P(c|x) = P(x|c)$ is equal to $P(c)/P(x)$, where $P(c|x)$ is probability of class $P(x|c)$ is probability of predictor class, $P(c)$ is the initial probability and $P(x)$ is the predictor probability. For this study, all reviews are divided into two classes which are for positives review and negative review.

Based on the results of the research by Mundalik (2018) found that Naive Bayes is the best-fit algorithms for the research project since the overall accuracy gained by Naive Bayes is 79.14% which is highest accuracy gained till the execution of this algorithm. Accuracy defines the high performance of model because this accuracy defines the model performance. In addition, precision gained by Naive Bayes is 0.62 which is directly contributes to the positive prediction for model (Mundalik, 2018).

2. Support Vector Machine

Sentiment analysis is treated as a classification task as it classifies the orientation of a text into either positive or negative. Support vector machine is broadly used in supervised machine learning method and analyzing of data is one of the major tasks in SVM. Support vector machine is one widely used in regression as well as in classification. It is used to classify the texts or reviews as positives or negatives in this study. According to the Zainuddin et al. (2014) stated that support vector machine is used to find different patterns in a dataset. SVM working is simple which is easy to analyze the border of decision and follows decision plane principal and classify class with maximum gap. This method works well with fewer samples (Zainuddin et al., 2014).

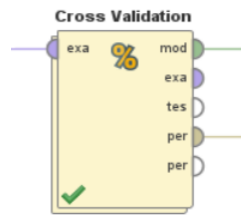
Based on the results of the research by Mundalik (2018) proved that support vector machine is performed well and the best-fit algorithms for the research project. Support vector machine achieved highest accuracy as 95.13% with the highest precision and recall as 0.99 and 0.99 respectively. This is high enough to call it as a best fit model (Mundalik, 2018). Support machine vector is one of the well-known and widely used this method because of its high accuracy results. The advantage of SVM is this method works well for text classification due to its advantages such as its potential to handle large features. Another advantage is SVM is robust when there is a sparse set of examples and also because most of the problem are linearly separable (Mullen et al., 2004).

3. Random Forest

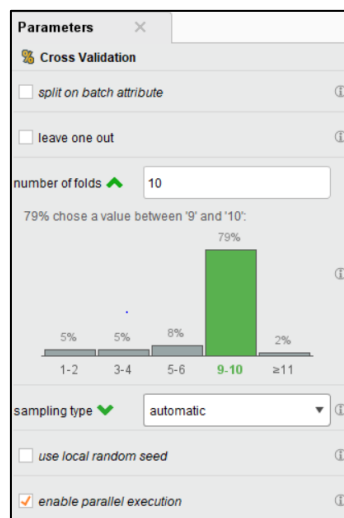
Random forest algorithm also known as random decision forest where it is one of the most used algorithms for sentiment analysis and mainly used for classification and regression task. Random forest is also a type of supervised machine learning algorithm based on ensemble learning. According to the Bahrawi (2019) stated that random forest is working is like ensemble approach because we must generate multiple trees while training of dataset and use these to deploy this method. Gupte et al. (2014) also said that because of selections of trees are random the correlation between them reduces and it influence to the high prediction power and efficiency.

Parmar et al. (2014) used random forest in movie reviews dataset sentiment analyzing because of its excellent performance and results that proved the random forest can also provide good and competitive results and can provide better results if hyper parameters are fine tuned. They further, used random forest and performed aspect-based sentiments analysis of movie review dataset. The first dataset resulted highest accuracy of 87.85% and for second dataset, it provided very promising results with accuracy of 91.00% (Parmar et al., 2014). Many researchers preferred random forest with another algorithm to perform ensemble approach.

5.2 Cross validation

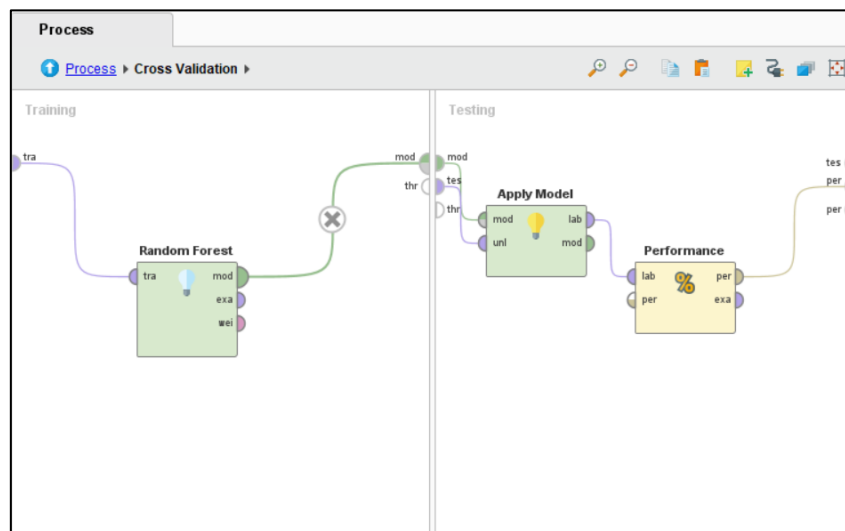


Cross validations operator performs a cross validation to estimate the statistical performance of a learning model. Based on the principles of data algorithm testing, it provides a better estimate of its performance. The samples used for training generally be divided into validation samples and training samples. Training samples are used to train the algorithm and validation samples are used as new data to evaluate the performance and operation of the algorithm.

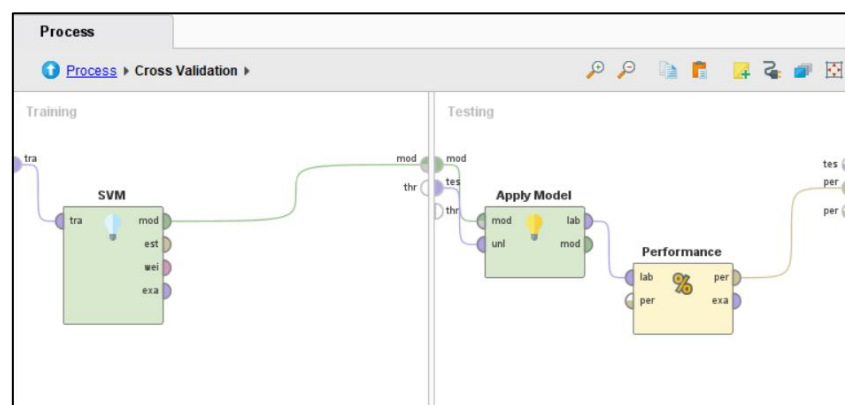


The parameter is set to 10 number of folds to estimate how accurately a model which learned by a particular learning operator that perform a practice. The cross-validation operator is a nested operator. It has two sub processes which are a training sub process and a testing sub process. The training sub process is used for a training a model. The trained model is then used in the testing sub process. The performance of the models which are naïve bayes, support vector machine and random forest are measured during the testing phase.

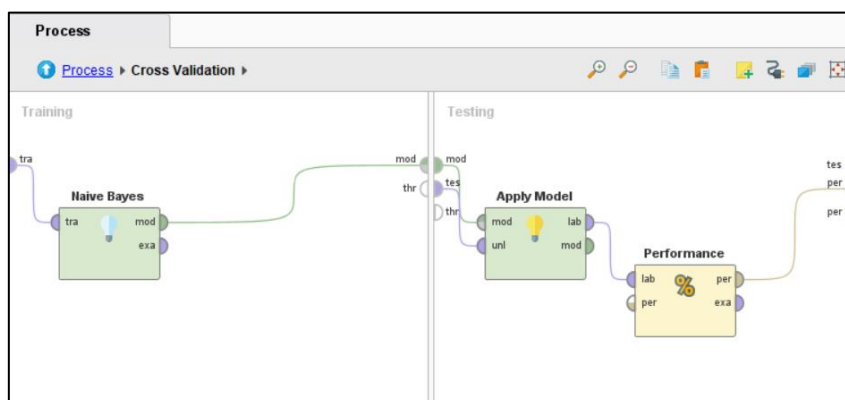
5.3 Modelling process in rapid miner



Applying Random Forest method for classification



Applying Support Vector Machine method for classification



Applying Naïve Bayes method for classification

5.4 Model comparison

Model comparison were made using confusion matrix. The confusion matrix contain value of how many cases did the model correctly predict and vice versa. Through this matrix, the models were compared based on its accuracy and its misclassification rate. Below are the confusion matrix of Naïve Bayes, Support Vector Machine and Random Forest model.

i. Naïve Bayes

accuracy: 88.21% +/- 2.37% (micro average: 88.21%)			
	true Positive	true Negative	class precision
pred. Positive	969	78	92.55%
pred. Negative	55	26	32.10%
class recall	94.63%	25.00%	

ii. Support Vector Machine

accuracy: 90.87% +/- 0.60% (micro average: 90.87%)			
	true Positive	true Negative	class precision
pred. Positive	1023	102	90.93%
pred. Negative	1	2	66.67%
class recall	99.90%	1.92%	

iii. Random Forest

accuracy: 90.78% +/- 0.46% (micro average: 90.78%)			
	true Positive	true Negative	class precision
pred. Positive	1024	104	90.78%
pred. Negative	0	0	0.00%
class recall	100.00%	0.00%	

Based on all the confusion matrix above, we can make side-by-side comparison as below:

Model	Accuracy	Misclassification rate
Naïve Bayes	88.21%	11.79%
Support Vector Machine	90.87%	9.13%
Random Forest	90.78%	9.21%

Interpretation: Based on the confusion matrix table, model support vector machine is the best model to detect polarity of emotions and opinion among user reviews of iPhone12. The Support Vector Machine (SVM) model has the highest accuracy of 90.87% with lowest misclassification rate of 9.13% compared to Naïve Bayes and Random Forest model.

6.0 FINDING

6.1 Preference of user to purchase Iphone12

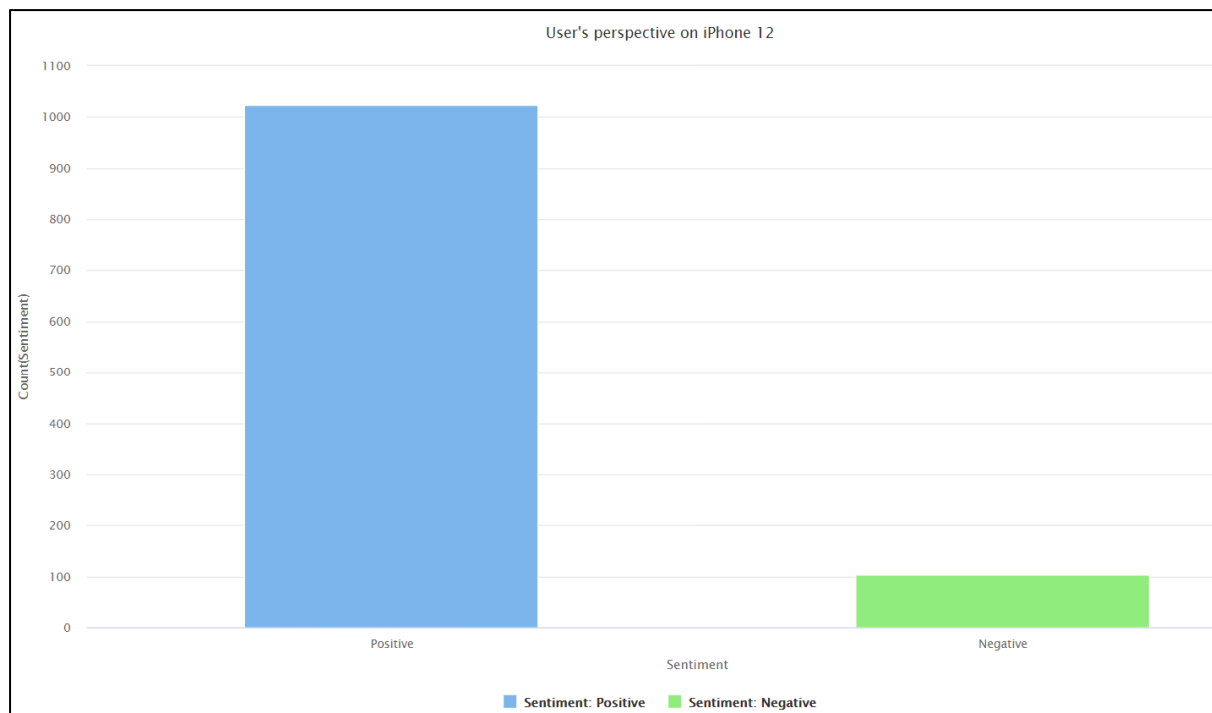


Figure of preference of user on iPhone12

After conducting a sentiment analysis, we determine whether the user's comments are positive or negative. From the bar chart above, we can see that most users have positive feedback about the iPhone 12. There are 1024 users who made positive comments and only 104 users who made negative comments. This shows that the probability of future sales performance for iPhone12 model will not drop since this iPhone12 model received more positive reviews among its users. Negative reviews may have been due to the fact that this time Apple is selling this phone without including the charger bricks in the box. Some users were frustrated by this situation as they have to buy separate charger bricks with additional cost included. Based on one of the comments by a user, he said, with a sarcastic tone, in regard of the situation where Apple did not provide the charger along with the purchase of the iPhone 12:

"Next time don't include phone also. And say u have one. With price of 100,000/-. Apple going mad. 🤔"

6.2 Rating of iPhone12 given by the user

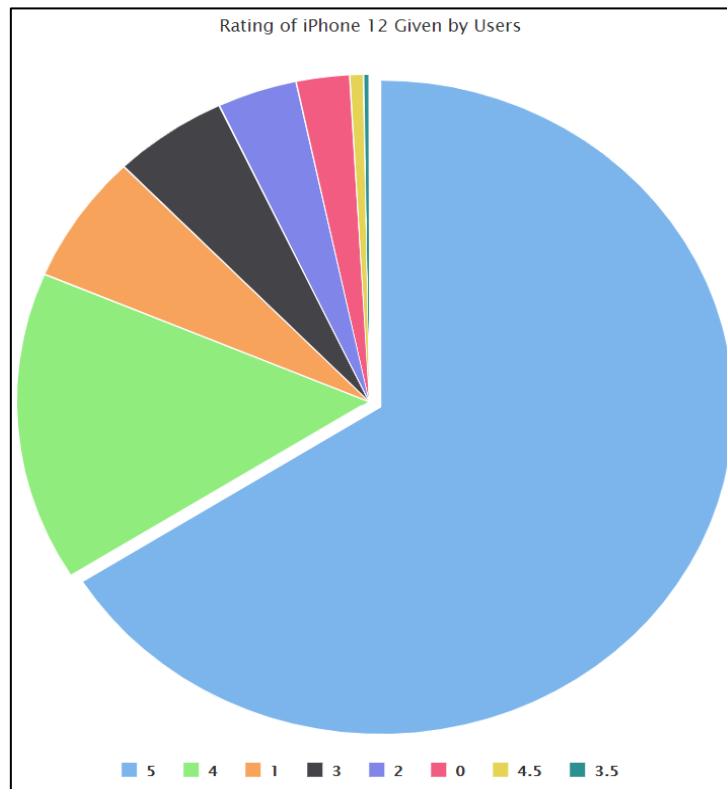


Figure for rating of iPhone12 given by the users

In general, most users of iPhone12 give 5-star rating to this phone model. This is not surprising, as a lot of users are making positive comments on the phone. However, there are only 755 users who give 5-star ratings. This shows that even though the users give a positive comment, it does not mean that they would give 5-star ratings. In addition, there are 175 users who gave 4-star ratings. There are only a few users who gave the rating below 4.

7.0 CONCLUSION

Research study on level sentiment analysis of iPhone 12 reviews has been conducted by using various natural language processing libraries in RapidMiner. Data has been tested on different supervised machine learning models. However, the support vector machine model gives highest accuracy of 90.87% on polarity detection of iPhone12 user reviews. This study is mainly divided into three parts which are by gathering online reviews, perform aspect level sentiment analysis and supervised developed machine learning models. The results of the sentiment analysis in this study show that the majority of users had positive expectations on the iPhone12, which led them to give high ratings. In addition, the feature that continues to be mentioned is its camera, display screen and battery performance. To the best of the student's knowledge, insufficient research has been carried out over the last few decades in a sentiment-based analysis of the iPhone domain. This research is useful in bridging the gaps. This research has brought significant contribution to the future buyers of iPhone12 as it has been proved that iPhone12 model is in general is a good phone model and worth to purchase according to its great feature quality that has improved compared to its previous model. It also makes a significant contribution to the Apple industry and to addressing its competitors. As stated in Chapter 2, all research objectives have been fulfilled.

REFERENCES

- Contributor, T. (2019, March 19). *sentiment analysis (opinion mining)*. SearchBusinessAnalytics.
<https://searchbusinessanalytics.techtarget.com/definition/opinion-mining-sentiment-mining>
- Data Mining Process. (2021). Oracle.
https://docs.oracle.com/cd/B19306_01/datamine.102/b14339/5dmtasks.htm
- Dey, L., Chakraborty, S., Biswas, A., Bose, B., & Tiwari, S. (2016). Sentiment Analysis of Review Datasets Using Naïve Bayes' and K-NN Classifier. *International Journal of Information Engineering and Electronic Business*, 8(4), 54–62.
<https://doi.org/10.5815/ijeeb.2016.04.07>
- Eckert, C., Katsikas, S. K., & Pernul, G. (Eds.). (2014). Differences between Android and iPhone users in their security and privacy awareness. *Lecture Notes in Computer Science*, 3–9. <https://doi.org/10.1007/978-3-319-09770-1>
- Everything There Is to Know about Sentiment Analysis*. (2020). MonkeyLearn.
<https://monkeylearn.com/sentiment-analysis/>
- Global Smartphone Market Share: By Quarter*. (2020, November 24). Counterpoint Research. <https://www.counterpointresearch.com/global-smartphone-share>
- Mickalowski, K., Mickelson, M., & Keltgen, J. (2008). Apple's iPhone launch: A case study in effective marketing. *The Business Review*, 9(2), 283-288.
- Probst, P., Wright, M. N., & Boulesteix, A. -. L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(3). <https://doi.org/10.1002/widm.1301>
- Statista. (2020, October 14). *Smartphone market in Malaysia - statistics and facts*. <https://www.statista.com/topics/6615/smartphones-in-malaysia/>
- Zainuddin, N., & Selamat, A. (2014). Sentiment analysis using Support Vector Machine. *2014 International Conference on Computer, Communications, and Control Technology (I4CT)*, 333–337. <https://doi.org/10.1109/i4ct.2014.6914200>

APPENDIX

Data

Filename	Description
IP12 combined	Data table for ip12 combined
iPhone12_Sentiment	Data table for sentiment analysis
iPhone12_weightcorrelation	Data table for weight correlation
iPhone12_wordlist	Data table for wordlist
iPhone12_wordvector	Data table for word vector

Model

Filename	Description
Naïve Bayes	Simple distribution for model Naive Bayes
Naïve_Bayes_Performance	Performance vector for model Naïve Bayes
Random forest	Simple distribution for model Random Forest
RandomForest_Performance	Performance vector for model Random Forest
SVM	Simple distribution for model SVM
SVM_Performance	Performance vector for model SVM

Process

Filename	Description
01_Text_processing_wordlist_wordvector	Pre-processing for wordlist and word vector
02_wordlist_frequency	Process for frequency of wordlist
03_wordvector_weightcorrelation	Process for weight by correlation of word vector
04_text_processing_sentiment_analysis	Pre-processing for sentiment analysis
05_Naive_Bayes_model_Crossvalidation	Process for Naïve Bayes classification
06_RandomForest_model_Crossvalidation	Process for Random Forest classification
07_SVM_model_Crossvalidation	Process for SVM classification