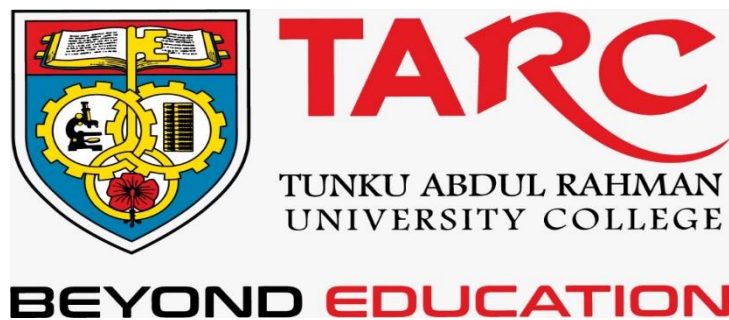


# **Sentiment analysis and visualisation of global COVID-19 vaccination plan using Naïve Bayes algorithm**

By

Tan Yi Hong



**FACULTY OF COMPUTING AND  
INFORMATION TECHNOLOGY**

**TUNKU ABDUL RAHMAN UNIVERSITY COLLEGE  
KUALA LUMPUR**

**ACADEMIC YEAR  
2020/21**

# Sentiment analysis and visualisation of global COVID-19 vaccination plan using Naïve Bayes algorithm

By

Tan Yi Hong

Supervisor: Dr. Lim Siew Mooi

A project report submitted to the  
Faculty of Computing and Information Technology  
in partial fulfillment of the requirement for the  
Bachelor of Computer Science (Honours)

**Department of Mathematical and Data Science**  
Faculty of Computing and Information Technology  
Tunku Abdul Rahman University College  
Kuala Lumpur

**Copyright by Tunku Abdul Rahman University College.**

All rights reserved. No part of this project documentation may be reproduced, stored in retrieval system, or transmitted in any form or by any means without prior permission of Tunku Abdul Rahman University College.

## Declaration

The project submitted herewith is a result of my own efforts in totality and in every aspect of the project works. All information that has been obtained from other sources had been fully acknowledged. I understand that any plagiarism, cheating or collusion or any sorts constitutes a breach of TAR University College rules and regulations and would be subjected to disciplinary actions.



---

Tan Yi Hong

Bachelor of Computer Science (Honours) in Data Science

ID: 20WMR08890

## Abstract

The COVID-19 pandemic has emerged as one of the world's most serious threats, and it is still very much a concern. Around the same time, we are in the middle of the world's most extensive vaccine program to fight against the deadly virus. Unfortunately, while the vaccine has given the battle against COVID-19 a new lease of life, it has also sparked a wave of anti-vaccine protests.

Therefore, it would be helpful to use sentiment analysis on recent Twitter data by using Twint API to crawl Tweets data about vaccination on Twitter to gauge public opinion on the COVID-19 vaccine. Therefore, this research project will carry out sentiment analysis using the Naïve Bayes approach and know the views of the people around the world towards their perception on the COVID-19 vaccine and determine the percentage of their responses that are positive, neutral, or negative emotions regarding the vaccine for the final classification.

## Acknowledgement

I want to thank the university, Tunku Abdul Rahman University College, for providing me with the opportunity to complete my final year project. This opportunity provides me with the chance to improve my Machine Learning and Natural Language Processing skills.

Next, I would like to express my gratitude to Dr. Lim Siew Mooi, my project supervisor, for inspiring me, encouraging guidance, and solid advice during the project. I will have a difficult time solving the issues I encountered without the direction of Dr. Lim Siew Mooi. In addition, Dr. Lim's skills and insights were invaluable to me throughout the project.

I also appreciate my project partner's willingness to share his knowledge and findings with me. Last but not least, I am grateful for my parents' support and encouragement in completing this project.

# Table of Contents

<b>Declaration.....</b>	<b>ii</b>
<b>Abstract.....</b>	<b>iii</b>
<b>Acknowledgement .....</b>	<b>iv</b>
<b>Table of Contents .....</b>	<b>v</b>
<b>List of Figures.....</b>	<b>vii</b>
<b>List of Tables .....</b>	<b>viii</b>
<b>1 Introduction .....</b>	<b>2</b>
1.1 Research Background .....	2
1.2 Problem Statement.....	2
1.3 Objectives .....	3
1.4 Research Scope .....	3
1.5 Research Contributions.....	3
1.6 Conclusion .....	4
<b>2 Literature Review.....</b>	<b>6</b>
2.1 Introduction.....	6
2.2 Sentiment Analysis .....	6
2.3 Data Visualization.....	7
2.4 COVID-19 Vaccination .....	10
2.5 Related Works.....	10
2.5.1 Sentiment Analysis of COVID-19 Tweets Using Deep Learning Models.....	10
2.5.2 Sentiment Analysis and its applications in fighting COVID-19 and infectious diseases .....	11
2.5.3 Sentiment Analysis of the Covid-19 Virus Infection in Indonesian Public Transportation on Twitter Data with Naïve Bayes and Decision Tree Approach .....	12
2.5.4 Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers.....	13
2.6 Classification approach.....	13
2.6.1 Naïve Bayes .....	13
2.6.2 Decision Tree.....	14
2.6.3 Random Forest.....	14
2.7 Summary .....	15
<b>3 Methodology and Requirements Analysis.....</b>	<b>18</b>
3.1 Methodology.....	18
3.1.1 The steps of sentiment analysis with Naïve Bayes approach:.....	18
3.2 Research Methods.....	20
3.2.1 Data collection .....	20
3.2.2 Text cleaning .....	21
3.2.3 Natural Language Processing (NLP) libraries .....	23
3.3 Chapter Summary & Evaluation.....	25
<b>4 Research Design.....</b>	<b>27</b>
4.1 Data Visualization Design .....	27
4.1.1 Types of visualisation for sentiment analysis .....	28
4.2 Model GUI Design.....	30
4.3 Chapter Summary and Evaluation .....	31
<b>5 Results .....</b>	<b>33</b>
5.1 Test Plan .....	33
5.1.1 Objectives of Experiment Testing .....	33

---

5.1.2	Objectives of System Testing .....	33
5.2	Experiment Testing.....	34
5.3	System Testing.....	37
5.4	Chapter Summary and Evaluation .....	41
<b>6</b>	<b>Discussions and Conclusion.....</b>	<b>43</b>
6.1	Summary .....	43
6.2	Achievements.....	43
6.3	Contributions .....	43
6.4	Limitations and Future Improvements.....	43
6.5	Issues and Solutions.....	44
<b>7</b>	<b>References .....</b>	<b>45</b>

## List of Figures

Figure 2.3-1: Example of Visualisation using Word Cloud to visualise the frequency of terms appeared in the dataset.....	8
Figure 3.1.1-1 : The major steps of a sentiment analysis.....	18
Figure 3.1-2 : Overall structure chart of the sentiment analysis with Naïve Bayes approach.....	20
Figure 4.1.1-1 : Bar chart showing the number of sentiments.....	28
Figure 4.1.1-2 : Pie chart showing the percentage of different sentiments in the data.....	29
Figure 4.1.1-3 : Word clouds that shows the frequency of words in the data.....	29
Figure 4.1.1-4 : Confusion matrix with 4 different combinations of predicted and actual values.....	30
Figure 4.2-1 : GUI of a sentiment detector.....	31
Figure 5.2-1 : Actual result of the word clouds for ‘Very Positive’ sentiment.....	35
Figure 5.2-2 : Actual result of the word clouds for ‘Positive’ sentiment.....	35
Figure 5.2-3 : Actual result of the word clouds for ‘Neutral’ sentiment.....	35
Figure 5.2-4 : Actual result of the word clouds for ‘Negative’ sentiment.....	36
Figure 5.2-5 : Actual result of the word clouds for ‘Very Negative’ sentiment.....	36
Figure 5.2-6 : Actual result of the bar chart for ‘Top 20 Unigrams words in Tweets’.....	36
Figure 5.2-7 : Actual result of the bar chart for ‘Top 20 Bigrams words in Tweets’.....	37
Figure 5.2-8 : Actual result of the bar chart for ‘Top 20 Trigrams words in Tweets’.....	37
Figure 5.3-1 : Actual result of Positive Sentiment Detector.....	39
Figure 5.3-2 : Actual result of Neutral Sentiment Detector.....	39
Figure 5.3-3 : Actual result of Negative Sentiment Detector.....	40
Figure 5.3-4 : Actual result of System Recoverability after pressing ‘Reset’ button.....	40



## List of Tables

Table 5.2-1 : Test plan for Experiment Testing.....	34
Table 5.3-1 : Test plan for System Testing.....	38

# Chapter 1

## **Introduction**

# 1 Introduction

## 1.1 Research Background

From 2019 until now, The COVID-19 pandemic has emerged as one of the world's most serious threats, and it is still very much a concern (J Glob Health, 2020). Around the same time, scientists worldwide have been trying hard to develop COVID-19 vaccines at a speed that we have never seen before for any other vaccine development.

They have successfully developed the COVID-19 vaccine to provide all people worldwide immunity to this deadly virus. As a result, many vaccines are available globally, such as Pfizer-BioNTech, Moderna, Johnson & Johnson / Janssen, AstraZeneca, CoronaVac, etc. While Malaysia so far has only five vaccines available: the Pfizer-BioNTech, CoronaVac, Sinovac, CanSino Biologics, and AstraZeneca vaccine (Tania J,2021).

In the meantime, we are in the middle of the world's most extensive vaccine program to fight against the deadly virus (JKJAV, 2021). While the vaccine has given the battle against COVID-19 a new lease of life, it has also sparked a wave of anti-vaccine protests (Nadirah H. R, 2021). Many people out there are afraid of the effectiveness and also the dangerousness of the COVID-19 vaccine. So, it would be helpful to use sentiment analysis on recent Twitter data crawled with Twint API to gauge public opinion on the COVID-19 vaccine to obtain people's feedback against the vaccination.

This research project will carry out sentiment analysis by using Naïve Bayes approach for all the people's opinions around the world towards their perception of the COVID-19 vaccine and determine the percentage of their responses that are positive, neutral, or negative emotions regarding the vaccine. After that, the entire progress flow and results will be visualised using Plotly, Bokeh, Word Clouds, etc.

As a result, this sentiment analysis model of COVID-19 vaccination and visualisation will benefit scientists and governments to obtain and gain people's opinion towards vaccination to take action for people to trust the impact of vaccination towards the entire world having campaigns.

## 1.2 Problem Statement

With the ongoing vaccination campaign globally, governments and scientists need to prove and show the success of the development of vaccines to strengthen the confidence of all people to fight against COVID-19. Without a doubt, leading the successful effect of vaccination and informative insights by visualisation will positively impact all people to be positive towards the vaccine and believe that the vaccine is the life-saving key in the battle against COVID-19.

It is crucial to let scientists and the government know more about the public opinions on the vaccination program to lower the rate of people starting the anti-vaccine protest. So, government and scientists will have to know the general view of the vaccine to encourage people to trust the vaccine and promote the positive side of the vaccine to change their minds. In such a way, more people can accept the vaccination to flatten the curve of the COVID-19 daily cases. Visualisation of this sentiment analysis model is a beneficial and

informative method to let people know the vaccination details by looking at the graphical chart, statistics, or tables.

### 1.3 Objectives

Throughout the project development phase, there are several objectives to be achieved, which are:

- Crawl the resources from Twitter with Twint API to collect people's tweets about the vaccine-related topic. The language will be specified to only English. The collected data will be processed and categorised into several classes: positive, negative, and neutral based on the text in all of the tweets, and find out their expression.
- To classify what people think of vaccines, either positive opinion such as they are confident, believes, and having the courage to the vaccination effect; or negative opinion such as they do not believe in it, and fear towards it.
- Develop a classification model with Naïve Bayes approach to applying in the real-world system that can manage to classify any sentences with the correct outcome.
- Visualise the entire analysis progress and results to graphical charts or dashboards using Bokeh, Plotly, or WordClouds to display valuable and informative insights.

### 1.4 Research Scope

The scope of this research project is to build and visualise a sentiment analysis model based on the COVID-19 vaccination. The dataset will be crawled from Twitter using Twint API, and from all of the sentences, the sentiment analysis model will classify different expressions of the people's opinions. The process and the result will be visualised using several visualisation tools such as Python, Matplotlib, Seaborn, Plotly, WordClouds, Bokeh etc. The model will then be developed and implemented into the real-world system with GUI by using TKinter from the Python GUI resources

### 1.5 Research Contributions

The contribution of this research project is to build a sentiment analysis model and visualise the progress and result to show the analysis outcome. This project will contribute as an essential source of better insights about the public opinions towards the vaccination of COVID-19 in the current situation. All the data will be crawled from Twitter, and visualise all the collected data for analysis purposes. The machine learning model will help classify the type of sentiment. The result will also be visualised for an overall image to everyone, especially government and scientists, to decide their following action to encourage people to be confident with the COVID-19 vaccination.

## 1.6 Conclusion

In a nutshell, this project will help researchers and reduce their workload from analysis through the internet. The sentiment analysis model will be automated to classify the sentiment from different classes for every sentence that is being collected. At the same time, the visualisation of the analysis progress and results will become an essential source of information that has valuable insights about the public opinions towards the COVID-19 vaccination from time to time. Other than that, the machine learning model was hoped to help society and hope that the COVID-19 pandemic can be cured successfully around the world.

# Chapter 2

## Literature Review

## 2 Literature Review

### 2.1 Introduction

Sentiment analysis is a method of analysing people's thoughts, sentiments, assessments, attitudes, and feelings based on what they write. According to some viewpoints, sentiment analysis determines what other people believe based on data such as written opinions. Therefore, the analysis was carried out on the written statement, according to the two views. People frequently express and post their ideas on social media due to the digital era's development, which makes us unable to avoid it.

When it comes to social media, Twitter is one of the most popular places to express themselves. This can be utilised as a data source for analysis. Because of its prominence, Twitter was chosen as a source of opinion mining. A concise explanation of a person's beliefs and thoughts can be up to 140 characters long on Twitter. This can be utilised as data in sentiment analysis to generate information and determine people's actual influence.

In this literature review, sentiment analysis and data visualisation details will be discussed after researching the open sources. We will get to know more about the information of COVID-19 vaccines, explanation and uses of different approaches of sentiment analysis, different types of data visualisation, and how to choose the suitable visualisation techniques.

### 2.2 Sentiment Analysis

Sentiment analysis is the process of analysing customer sentiment utilising natural language processing, text analysis, and statistics. The finest companies know their consumers' feelings—what they're saying, how they're expressing it, and what they mean (Algorithmia, 2018). Advances in deep learning, like many other domains, have pushed sentiment analysis to the forefront of cutting-edge algorithms. They have been using natural language processing, statistics, and text analysis to extract and categorise the sentiment into positive, negative, or neutral categories.

Sentiment analysis nowadays has been widely implemented for the use of brand monitoring. One of the most well-known applications of sentiment analysis is to obtain a complete 360-degree perspective of how customers and stakeholders perceive your brand, product, or company. Product reviews and social media, for example, are widely available media that can give crucial insights into what your firm is doing correctly or poorly. Companies can utilise sentiment analysis to assess the impact of a new product, ad campaign, or a customer's reaction to recent company news on social media.

Sentiment analysis was also implemented for the use of customer service. It is frequently used by customer service employees to automatically categorise incoming user email into "urgent" or "not urgent" categories based on the email's sentiment, proactively detecting unhappy users. The agent then prioritises fixing the users with the most pressing issues first. Understanding the view and intent of a specific case becomes

increasingly critical as customer support becomes increasingly automated through machine learning.

Other than that, sentiment analysis has also been used for market research and analysis. Sentiment analysis is a technique used in business intelligence to determine the subjective reasons why customers respond to something. (For example, why do customers buy a product? What are their thoughts on the user interface? Did the level of customer service fulfil their expectations?). Sentiment analysis can be used to examine trends, ideological bias, opinions, assess reactions, and more in political science, sociology, and psychology.

In this project, sentiment analysis is carried out to analyse the opinions of the people around the world towards their perception of the COVID-19 vaccine from Twitter using Twint API and classify the percentage of their responses that are positive, neutral, or negative emotions regarding the vaccine using the Naïve Bayes approach.

## 2.3 Data Visualization

Text data analysis is becoming increasingly simple. For text data analysis, popular computer languages like Python and R have excellent packages. Unfortunately, people believed that you needed to be a coding specialist to complete these kinds of challenging tasks. However, with the more evolved and better versions of libraries, performing text data analysis with only essential and beginner-level coding skills has become more accessible (Rashinda N S, 2021).

The graphical depiction of information and data is known as data visualisation. Data visualisation tools make it easy to examine and comprehend trends, outliers, and patterns in data by employing visual elements like charts, graphs, and maps. Data visualisation tools and technologies are critical in the Big Data environment to analyse enormous volumes of data and make data-driven decisions.

Colors and patterns attract our attention. For example, we can rapidly distinguish between red and blue, and a square from a circle. Everything in our culture is visual, from art and marketing to television and movies. Another type of visual art that piques our curiosity and keeps our gaze fixed on the message is data visualisation. We can rapidly spot trends and outliers while looking at a chart. If we can see something, we immediately assimilate it. It's purposeful storytelling. For example, figure 2.3-1 below shows the visualisation of terms frequencies using the Word Cloud. It illustrates the most frequently appeared word in the form of graphical representation.



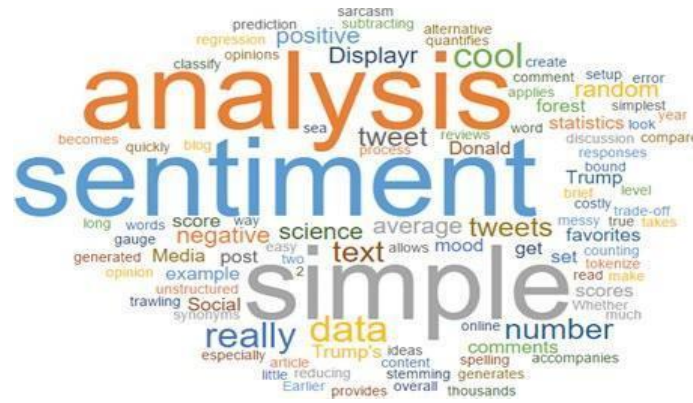


Figure 2.3-1: Example of Visualisation using Word Cloud to visualise the frequency of terms appeared in the dataset.

There are many other types of visualisations available in the public community creations. However, simple bar graphs or pie charts are typically the first things that come to mind when thinking of data visualisation. While these are an essential aspect of data visualisation and a frequent starting point for many data visualisations, the proper visualisation must be combined with the appropriate set of data.

**Common general types of data visualisations include:**

- Charts - Frequently used to make enormous amounts of data and the links between them easier to comprehend.
- Tables - Used to organise too detailed or complex data to be fully represented in the text, allowing the reader to see the results immediately.
- Graphs - A popular method of visually showing data relationships.
- Maps - Utilised to analyse, display, and show geographically connected data in the form of maps.
- Infographics - In the form of a graph or an image, showcase vast amounts of data and information.
- Dashboards - A data visualisation tool that enables all users to comprehend the critical analytics to their company, department, or project.

**More specific examples of methods of data visualisations:**

- Area Chart - It's a systematic way to display data that shows a time-series relationship.
- Bar Chart - Shows categorical data as rectangular bars with lengths and heights that match each data point's values.
- Box-and-whisker Plots - a method of presenting a set of data on an interval scale in a graphical format.
- Bubble Cloud - A form of a chart that, like a scatter chart, displays data by position on an axis and adds a third variable in the size of the bubble.
- Bullet Graph - Used to relate qualitative ranges by comparing one value, represented by a horizontal bar, to another value, represented by a vertical line.
- Cartogram - As an isodemographic map, it represents an area in the context of the value of a variable linked with it.

- Circle View - Depicts the relationship between the pieces of something and the total.
- Dot Distribution Map - A map in which the presence of a feature or phenomena is shown by the density of dot symbols of the same size.
- Gantt Chart - A time-based visual representation of tasks
- Heat Map - A graphical representation of visitor behaviour data as hot and cold patches using a warm-to-cool colour scheme
- Highlight Table - To use colour to compare categorical data
- Histogram - Gives a visual depiction of how data is distributed.
- Matrix - A project management and planning tool for analysing and understanding data set interactions.
- Network - An approach for visualising intricate interactions between several aspects
- Polar Area - Identical to a traditional pie chart, only the sectors have all the same angle and change only in how far each sector extends from the circle's centre.
- Radial Tree - A technique of showing a tree structure that spreads outwards in a radial fashion.
- Scatter Plot (2D or 3D) - To notice and demonstrate correlations between two numeric variables
- Streamgraph - Ideal for showing large datasets and discovering trends and patterns over time in various categories.
- Text Tables - Used to organise information that is too complex or detailed to be fully described in the text
- Timeline – To display a series of events or a process that occurred over time.
- Treemap - To represent a big quantity of data in a hierarchical, tree-structured diagram with rectangle sizes sorted from most significant to lowest
- Wedge Stack Graph - A visualisation method that uses a radial approach to represent hierarchical data.
- Word Cloud - A word visualisation that ranks the most frequently used terms in a text from tiny to large, based on how often they appear.

In this project, the visualisation of the entire sentiment analysis progress and results will become an essential source of information that has valuable insights about the public opinions towards the COVID-19 vaccination from time to time. The type of data visualisation used for this sentiment analysis is the graphical charts or dashboard using Bokeh, Plotly, or Word Cloud to display helpful and informative insights about the study.

## 2.4 COVID-19 Vaccination

The first global rollout of the COVID-19 vaccination occurred in the UK. On 8 December 2020, a 90 years old British woman named Margaret Keenan had received the first-ever COVID-19 vaccine shot in the world (Heidi L, 2021). The first-ever COVID-19 vaccine shot was the Pfizer/BioNTech vaccine (BBC, 2020). They had prepared the first 800,000 doses of Pfizer/BioNTech vaccine to be vaccinated for the over-80s people and health care staff in their country. More than 2.12 billion people have already been vaccinated with vaccines, such as Moderna, Johnson & Johnson / Janssen, AstraZeneca, CoronaVac, SinoVac, etc (OurWorldInData, 2021).

It's been a long year of sickness, devastation, grief, and despair, but the global rollout of COVID-19 vaccines (Donato P M, 2021) has brought relief and renewed hope to many. However, the debate about vaccination advancements, accessibility, efficacy, and side effects continue, and it permeates news stories, and daily Twitter feeds. Our internet exposure, on the other hand, is limited to our echo chambers. As a result, the goal of this project is to broaden our view on the global epidemic by utilising the power of Twitter text data.

It would be almost impossible for a human to read and comprehend everything said about COVID-19 vaccinations on Twitter. Fortunately, using textual fertilisation, sentiment analysis, and word cloud visualisations, we can look into an incredibly complicated and wide-ranging dialogue using natural language processing (NLP) approaches.

## 2.5 Related Works

### 2.5.1 Sentiment Analysis of COVID-19 Tweets Using Deep Learning Models

The new coronavirus disease (COVID-19) is an ongoing pandemic that has sparked widespread concern worldwide. Spreading misleading information on social media platforms like Twitter, on the other hand, is exacerbating the disease's situation. The goal of this research is to examine Indian netizens' tweets during the COVID-19 lockdown. The tweets were collected between March 23, 2020, and July 15, 2020, and the text was labelled as fear, sadness, rage, and joy (Chintalapudi N, 2021).

This study used data from Indian user tweets from the Twitter website during the COVID-19 shutdown period in-country for investigation. The data collection, which contains 3090 tweets, was retrieved from GitHub and included clean tweets on COVID-19, coronavirus, and lockdown topics. For the analysis, a dataset of extracted tweets from the Indian Twitter network was used. COVID-19, coronavirus, and lockdown were among the subjects discussed in the tweets.

The sentiment was analysed using the Bidirectional Encoder Representations from Transformers (BERT) model, a new deep-learning model for text analysis

and performance that was compared to three other models: support vector machines (SVM), logistic regression (LR), and long-short term memory (LSTM).

Every sentiment's accuracy was calculated separately. The BERT model had an accuracy of 89 percent, whereas the other three models had an accuracy of 74.75 percent, 75 percent, and 65 percent, respectively. The accuracy of each sentiment categorisation ranges from 75.88 to 87.33 percent, with a median accuracy of 79.34 percent, which is a significant value in text mining algorithms.

According to these findings, during COVID-19, Indian tweets had a high prevalence of keywords and related terms. Furthermore, this work clarifies public perceptions of pandemics and guides public health authorities toward a more prosperous society.

### **2.5.2 Sentiment Analysis and its applications in fighting COVID-19 and infectious diseases**

In December 2019, the COVID-19 pandemic, caused by the new coronavirus SARS-CoV-2, struck China unexpectedly. According to the World Health Organization, tens of millions of identified cases and thousands of verified deaths have been reported globally. The infection is making its way through social media websites.

As a result, throughout numerous outbreak-related occurrences, these social media channels are experiencing and conveying various views, ideas, and feelings. Thus, big data is a valuable tool for computer scientists and researchers in studying people's reactions to current events, particularly those related to the outbreak.

There has never been a study that looked at multiple diseases using sentiment analysis (Alamoodi, 2021). As a result, the goal of this study was to review and analyse articles published in the last ten years about the occurrence of various infectious diseases, such as epidemics, pandemics, viruses, and outbreaks, to understand the application of sentiment analysis better and obtain the most essential literature findings. From January 1, 2010, to June 30, 2020, articles on related themes were carefully searched in five central databases: PubMed, ScienceDirect, Scopus, IEEE Xplore, and Web of Science. These indexes were deemed comprehensive and trustworthy enough to cover the research area.

There were a total of  $n = 28$  journals chosen for the systematic review based on our inclusion and exclusion criteria. These articles were combined into a unified taxonomy to describe the present literature's relevant current viewpoints in four primary categories: lexicon-based models, machine learning-based models, hybrid-based models, and persons. In addition, the publications were categorised into three categories: illness mitigation motivations, data analysis

motives, data, social media platform, and community challenges faced by researchers.

Other parts of the systematic review were included in the study, such as the methodology followed by the periodic review and demographic information of the literature distribution. As a result, interesting patterns in the literature were discovered, and the articles were classified accordingly. In addition, the existing state of knowledge and research prospects in this topic was highlighted in this study, which encouraged more efforts to comprehend this research topic better.

### **2.5.3 Sentiment Analysis of the Covid-19 Virus Infection in Indonesian Public Transportation on Twitter Data with Naïve Bayes and Decision Tree Approach**

(C. Sari and Y. Ruldeviyani, 2020) The Covid-19 virus, which first appeared in early 2020, became a terrifying pandemic for the entire world, including Indonesia. Because the Covid-19 virus may be transmitted through human contact, infection was quick. This situation is causing concern in society. Furthermore, similar problems are shared by passengers on public transit, particularly commuter lines. If commuter line passengers spread the Covid-19 virus to the commuter line in huge numbers and push each other, it will be a source of concern.

Many passengers use Twitter to express their thoughts on the spread of the Covid-19 epidemic. As a result, numerous viewpoints emerge, which can be positive, negative, or even neutral. As a result, a study was conducted to examine the sentiment of the Covid-19 transmission to commuter line passengers to see what they thought. A comparison of two approaches was used in this study, and Nave Bayes surpassed the Decision Tree with an accuracy of 73.59 percent. Furthermore, when compared to the other two classifications, the sentiment analysis resulted in a positive category.

Text mining, a process of extracting essential data from content, is used in this sentiment analysis study. Text mining is a type of data mining that entails document preparation. Document pre-processing converts unstructured text into structured data. After that, the structured data is categorised using a data mining classification algorithm. In this study, the Naïve Bayes and Decision Tree methods were used. The Naïve Bayes method was chosen because of its solid assumptions and great accuracy. While the Decision Tree method was selected because it is more straightforward and straightforward to utilise.

## 2.5.4 Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers

The World Health Organization (WHO) declared COVID-19, also known as CoronaVirus Disease of 2019, as a pandemic on March 11, 2020. Unprecedented demands have mounted on each government to impose stringent requirements for population control by analysing cases and properly deploying available resources. People are experiencing worry, dread, and worry due to the increasing number of exponential instances worldwide. The world population's mental and physical health has been directly proportional to this pandemic sickness.

(Chakraborty K, 2020) This study intends to highlight that tweets featuring all COVID-19 and WHO handles have failed to guide people through the pandemic threat effectively. This research looks at two categories of tweets collected during the pandemic. In one scenario, over 23,000 of the most retweeted tweets from January 1, 2019 to March 23, 2020 were studied, and it was discovered that the majority of the tweets expressed neutral or negative opinions. On the other hand, a dataset of 226,668 tweets recorded between December 2019 and May 2020 was evaluated, revealing that netizens tweeted the most positive and neutral tweets.

The study indicated that netizens were preoccupied with retweeting the negative tweets even though most individuals tweeted positive things about COVID-19. Furthermore, no valuable words could be located in Word Cloud or computations based on word frequency in tweets. The assertions were supported by a suggested model that used deep learning classifiers to achieve allowable accuracy of up to 81 percent. The authors also propose using a fuzzy rule basis based on a Gaussian membership function to identify sentiments in tweets correctly. The accuracy of the model above is up to a maximum of 79 percent.

## 2.6 Classification approach

### 2.6.1 Naïve Bayes

Naïve Bayes is a grouping model that uses the division of words in a document to determine probability in each class. This method has been tested with numerous other algorithms and has a high level of accuracy when used for sentiment analysis. The following is the theory that will be utilised to forecast probabilities:

$$P(features) = \frac{P(label) * P(features|label)}{P(features)}$$

$P(label)$  is the label's previous probability. The prior probability that is classed as a label is  $P(features|label)$ . The prior probability that occurred is  $P(features)$ .

Because Naïve was given the assumption that all features are free, the formula might be recast as:

$$P(features) = \frac{P(label) * P(label) * ... * P(fn|label)}{P(features)}$$

Text classification and spam filtering are two areas where the Naive Bayes Model excels. Working with the NB algorithm has the following advantages:

- To learn the parameters, only a tiny amount of training data is required.
- When compared to more sophisticated models, they can be taught relatively quickly.
- The following are the primary drawbacks of the NB Algorithm:
- It's a good classifier but a terrible estimator.
- It works well with discrete numbers but not with serial numbers.

### 2.6.2 Decision Tree

The decision tree approach is a model of a tree or tree that is used to predict test results. Attributes act as both a node and a branch in a tree. The root is the highest node in a decision tree. The classification process starts at the tree's root node. The phases of the decision tree algorithm begin with:

- 1.Prepare the data for training.
- 2.Choose the attributes as root using Information Gain (ID3).
- 3.For each value, make a branch.
- 4.Repeat for each branch until all cases in that branch have the same class.

Because decision trees are supervised algorithms, they must be trained using annotated data. As a result, the central notion is the same as any text classification: given a set of documents (for example, TFIDF vectors) and their labels, the algorithm will determine how much each word connects with each label.

It may be discovered, for example, that the term "great" frequently appears in positive texts, but the term "awful" frequently appears in negative papers. It creates a model that can assign a label to any document by combining all of these data.

### 2.6.3 Random Forest

Random Forest algorithm is a supervised classification algorithm. It's an ensemble learning method based on the decision tree algorithm. Ensemble classification methods are learning algorithms that create a group of classifiers rather than a single classifier and then classify new data points by voting on their predictions.

Ensemble learning is a sort of learning in which numerous versions of the same algorithm are combined to produce a more effective prediction model. For example, the random forest algorithm combines several methods of the same sort, such as numerous decision trees, to create a forest of trees, hence the name "Random Forest." Both regression and classification jobs can benefit from the random forest approach.

To improve robustness over an individual estimator, this Ensemble approach aggregates the predictions of some base estimators built with the decision tree approach. Random Forest produces a large number of classification trees, which is referred to as a forest. If we wish to classify new data, each tree offers one vote to its category forecast. The forest selects the category with the most votes. The more trees in the random forest, the higher the accuracy of the results.

With replacement, a new training data set is constructed from the original data set. Then, using random feature selection, a tree is grown. Pruning is not done on mature trees. Random Forest's precision is unrivalled thanks to this method. Random Forest is similarly quick, resistant to overfitting, and allows users to create as many trees as they wish.

The phases of Random Forest algorithm is as follows (Bahrawi B, 2019):

1. Using the original data, create  $n_{tree}$  bootstrap samples.
2. Grow an unpruned classification or regression tree for each bootstrap sample, with the following modification. At each node, instead of choosing the best split among all predictors, randomly choose  $m_{try}$  of the predictors and choose the best split among those variables. (Bagging can be viewed as a specific example of random forests where the number of predictors is  $m_{try} = p$ .)
3. By combining the forecasts from the  $n_{tree}$  trees, predict new data.

(Muhammad A F, 2018) Random Forests have grown in prominence in recent years due to their superior performance in classification tasks in fields such as bioinformatics and computational biology. Other works also use Random Forest for text classification, such as hate speech identification and authorship profiling.

## 2.7 Summary



This chapter discusses the literature review for sentiment analysis and data visualisation in detail after the research. Several related works about sentiment analysis related to COVID-19 have also been discussed about their way of doing their projects. We get to know the use and benefits of sentiment analysis, and different types of data visualisation and its application, and the knowledge of COVID-19 vaccination in the global. Lastly, several machine learning techniques can classify sentiments such as Naïve Bayes, Decision Tree, and Random Forest. All of them have different methods to classify and predict the results.

# Chapter 3

## **Methodology and Requirements Analysis**

## 3 Methodology and Requirements Analysis

### 3.1 Methodology

#### Machine Learning Life Cycle

Machine learning has given computers the ability to learn on their own without having to be explicitly programmed. However, how does a machine learning system function? As a result, the machine learning life cycle can be used to explain it. A machine learning project's life cycle is a cyclic method for developing a practical machine learning project. The life cycle's primary goal is to find a solution to the problem or project.

Understanding the problem and knowing the problem's purpose are the most critical aspects of the entire procedure. As a result, we must first comprehend the problem before beginning the life cycle, as a positive outcome is contingent on a thorough comprehension of the situation. To address an issue, we develop a machine learning system called a "model" in the complete life cycle process, and this model is built by providing "training." However, we need data to train a model. Therefore the life cycle begins with data collection.

#### 3.1.1 The steps of sentiment analysis with Naïve Bayes approach:

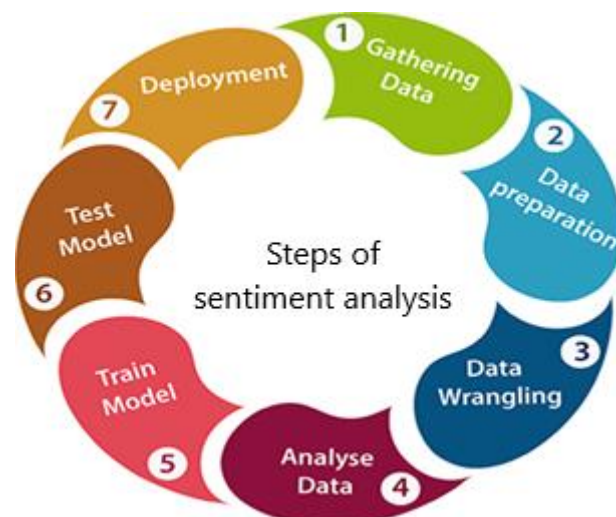


Figure 3.1.1-1 : The major steps of a sentiment analysis.

#### Step 1 - Gathering Data

The goal of this step is to identify and obtain all data-related problems. It is to collect data, and integrate data obtained from Twint API/Twitter API to get a coherent set of data called as a dataset.

#### Step 2 - Data preparation

This step is to put the data collected into a suitable place and prepare it in our sentiment analysis model training. Thus, the first process is data exploration followed by data pre-processing and stored into dataframe.

**Step 3 - Data wrangling**

After processing the data collected, this step helps clean and convert raw data into a usable format such as removing stopwords from texts and lemmatising the sentences. This step is essential to clean the data issues, such as invalid data and unreadable text.

**Step 4 - Analyse data**

After cleaning data, this step is to build a Naïve Bayes model to analyse the data using various analytical techniques and review the outcome of the sentiment analysis. This is to take the data and use machine learning algorithms to build the model.

**Step 5 - Train the model**

This step is to train the model to improve its performance for better outcome of the sentiment result.

**Step 6 - Test the model**

In this step, the model is checked for sentiment accuracy by providing a test dataset to it.

**Step 7 - Deployment**

After the above steps is complete, the final Naïve Bayes model is deployed in the real-world system with GUI using Tkinter from Python GUI resources.

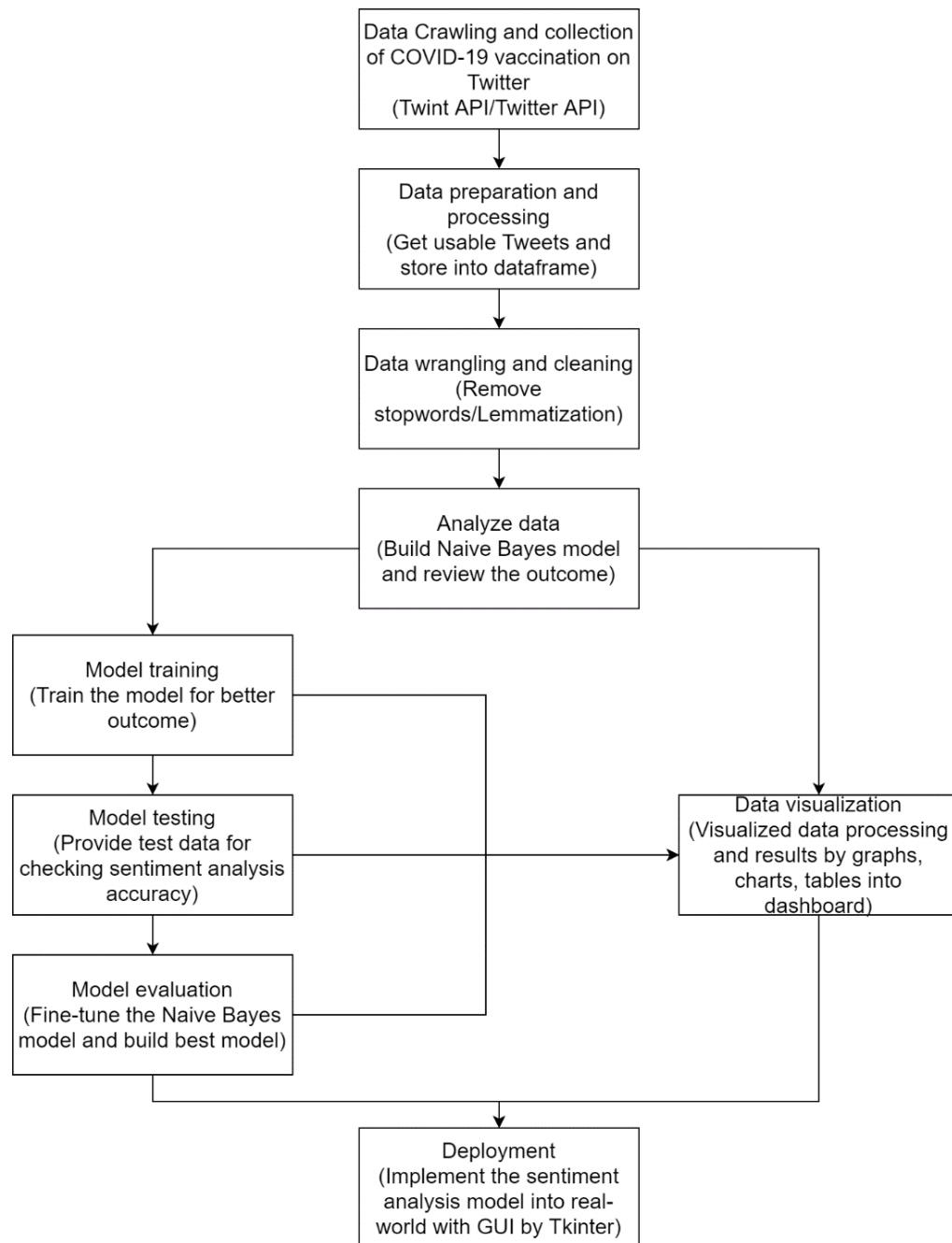


Figure 3.1-2 : Overall structure chart of the sentiment analysis with Naïve Bayes approach.

## 3.2 Research Methods

### 3.2.1 Data collection

The procedure of collecting, measuring, and evaluating correct research insights using established procedures is referred to as data collection. Based on the facts gathered, a researcher might evaluate their hypothesis. Regardless of the subject of

study, data collecting is usually the first and most significant phase in the research process. Depending on the information needed, different approaches to data gathering are used in various disciplines of study.

The most important goal of data collecting is to collect information-rich and accurate data for statistical analysis to make data-driven research decisions. In addition, data must be collected and kept in a form that makes sense for the business problem to use to develop viable artificial intelligence (AI) and machine learning (ML) solutions.

Data collection allows us to keep track of past events to utilise data analysis to uncover repeating patterns. Using machine learning algorithms, you may create predictive models that look for trends and forecast future changes based on those patterns.

Because predictive models are only as strong as the data they're built on, reasonable data collecting procedures are essential for creating high-performing models. The data must be devoid of errors (garbage in, garbage out) and contain relevant information to work at hand. A debt default model, for example, would not profit from tiger population sizes but would benefit over time from gas costs.

In this project, the data collected in the Tweets related to 'COVID-19 Vaccination' topics from Twitter. The tool that use to crawl the related Tweets is the Twint API. Twint is a Python-based advanced Twitter scraping application that allows you to scrape Tweets from Twitter profiles without using Twitter's API.

Twint makes use of Twitter's search operators to allow you to scrape Tweets from specific individuals, scrape Tweets referring to particular themes, hashtags, and trends, and sort out sensitive information like email and phone numbers from Tweets. Twint also creates unique Twitter queries that allow you to scrape a Twitter user's followers, Tweets they've liked, and who they follow without having to utilise any login, API, Selenium, or browser emulation.

There are several benefits of using Twint as a web crawler. First, it can crawl almost all Tweets without limitation, as Twitter API only limits the user to crawl a maximum of 3200 Tweets. Second, Twint also provides a fast initial set-up, and it can be used without the need to perform authentication of Twitter.

### 3.2.2 Text cleaning

Natural Language Processing is being used to generate a variety of activities and goods. Any form of a model, such as classification, Q&A model, sentiment analysis, and others, uses text as its primary input.

As an example, assume that the text contains various symbols and words that do not transmit meaning to the model during training. As a result, we'll delete them before feeding the model efficiently. Data Preprocessing is the term for this procedure. It's also referred to as Text Cleaning.

Text is a type of data that has been around for millennia throughout human history. The importance of data available in the form of text is defined by all sacred texts influencing all religions, all poets' and authors' compositions. All scientific explanations by the brightest minds of their times, all political documents defining our history and future, and all kinds of explicit human communication.

Text is nothing more than a string of words, or more specifically, a string of letters. However, when dealing with language modelling or natural language processing, we are usually more concerned with the words as a whole rather than just the character-level depth of our text data. One explanation for this is that individual characters in language models don't have much "context." For example, characters like 'd,' 'r,' 'a,' and 'e' have no meaning on their own, but when rearranged into a word, they produce the word "read," which could describe what you're presumably doing right now.

The removal of undesirable characters is the first step in the text cleaning process. For example, if we crawled some text from HTML/XML sources, we'll have to remove all tags, HTML entities, punctuation, non-alphabets, and any other characters that aren't part of the language. Regular expressions, which may be used to filter out most undesired texts, are the most common techniques of such cleaning.

Important English characters such as full stops, question marks, and exclamation symbols are kept in some systems. Consider the following scenario: you want to perform sentiment analysis on human-generated tweets and categorise them as extremely angry, furious, neutral, pleased, and very happy. Because there are some instances that only words can convey, simple sentiment analysis may struggle to distinguish between a happy and a joyful mood. "The drink is nice.", and "The. Drink. Is. Nice!!!!!!!!!" have different sentiments. The overuse of the exclamation marks shows an "extra" feeling and is different from the calm sentences.

Emoticons, which are non-alphabetic characters, also play a part in sentiment analysis. ":), :(, --, :D, xD," all of these, when properly analysed, can aid in sentiment analysis. Even if you're trying to build a system that can categorise whether a statement is sarcastic or not, such minor features can help.

The dataset's text cleaning will start from the Tokenisation and Capitalisation or De-capitalization of words from the sentence. For example, "I feel grateful after this!" will be tokenised into "i", "feel", "grateful", "after", "this", "!", ". The next step

after this pre-processing is to remove stopwords from the data. Stopwords are words that are used so frequently that they have lost their semantic significance. Stopwords include words like "of," "are," "the," "it," and "is." Getting rid of stopwords can be a good idea in applications like document search engines and document classification, where keywords are more significant than broad terms.

Next, lemmatisation and stemming will be performed to the text data. Both stemming and lemmatisation have the purpose of reducing a word's inflectional and occasionally derivationally related forms to a common base form. As a result, the stemming and lemmatising step helps reduce the number of overall terms to a few "root" terms. For example, Organizer, organises, organisation, organised all these words will be reduced to their root term, most probably becoming "organiz".

Stemming is a rudimentary method of reducing phrases to their base by simply setting rules for slicing off some characters at the end of the word, which, in most cases, produces good results. While Lemmatization is a more organised way of achieving the same thing as stemming, it also requires some vocabulary and morphological analysis.

Because affixes of words include additional information that can be used, stemming and lemmatisation should only be done when necessary. For example, the terms "faster" and "fastest" share the same root but have different semantic meanings.

So, if the application relates to the word, as most search engines and document clustering systems do, stemming and lemmatisation process may be an alternative, but stemming and lemmatisation process may be eliminated for applications that require some semantic analysis.

### 3.2.3 Natural Language Processing (NLP) libraries

#### Natural Language Toolkit (NLTK)

In Python, NLTK is a helpful package that helps with classifications, stemming, tagging, parsing, semantic reasoning, and tokenisation. It's the most essential tool in natural language processing and machine learning. It now serves as a basis for Python developers who are just getting their feet wet in the industry (and machine learning).

Steven Bird and Edward Loper of the University of Pennsylvania created the library essential in groundbreaking NLP research. NLTK, Python libraries, and other technologies are now used in many university courses throughout the world.

(Link : <https://www.nltk.org/>)



### **TextBlob**

TextBlob is a must-have for Python developers who are just getting started with NLP and want to get the most out of their first experience with NLTK. It essentially gives newcomers an easy-to-use interface to assist them in learning the most basic NLP tasks, such as sentiment analysis, pos-tagging, and noun phrase extraction. Thus, it comes in handy while creating prototypes. However, it also inherited NLTK's major flaws: it's simply too slow to assist developers dealing with NLP Python production use rigours.

(Link : <https://textblob.readthedocs.io/en/dev/>)

### **CoreNLP**

This Java library was created at Stanford University and is available for download. It does, however, come with wrappers for a variety of languages, including Python. That's why it's handy for Python developers who want to try their hand at natural language processing. In addition, the library is speedy and well-suited for use in product development environments. Furthermore, some CoreNLP components can be combined with NLTK, increasing the latter's efficiency.

(Link : <https://stanfordnlp.github.io/CoreNLP/>)

### **Gensim**

Gensim is a Python package that uses vector space modelling and a topic modelling toolkit to find semantic similarities between two documents. With the help of efficient data streaming and incremental algorithms, it can handle big text corpora, which is more than we can say for competing packages that solely target batch and in-memory processing. Its excellent memory use optimisation and processing speed are what we adore about it. These were accomplished with the help of NumPy, a Python module. The vector space modelling capabilities of the programme are also excellent.

(Link : <https://github.com/RaRe-Technologies/gensim>)

### **spaCy**

spaCy is a new library that was created with production in mind. That is why it is far more user-friendly than competing Python NLP packages such as NLTK. spaCy has the quickest syntactic parser on the market right now. Furthermore, because the toolkit is developed in Python, it is speedy and efficient. Compared to the other libraries we've looked at so far, spaCy supports the fewest languages (seven). However, given the expanding prominence of machine learning, natural language processing, and spaCy as a crucial library, the tool may soon support more programming languages.

(Link : <https://spacy.io/>)

### **Polyglot**

This little-known library provides a wide variety of analyses as well as extensive language coverage. It also works very quickly, thanks to NumPy. Polyglot is similar to spaCy in that it is very efficient, simple, and an ideal solution for applications that require a language that spaCy does not support. The library stands out from the crowd since it uses pipeline methods to request the use of a specific command in the command line.

(Link : <https://polyglot.readthedocs.io/en/latest/index.html>)

### **scikit-learn**

This valid NLP package gives programmers access to a variety of algorithms for creating machine learning models. It has a lot of functionality for dealing with text categorisation problems utilising the bag-of-words method of building features. The intuitive classes methods are the library's strength. Additionally, scikit-learn comes with good documentation to assist developers in making the most of its features. The library, on the other hand, does not employ neural networks for text pre-processing. Therefore, it is preferable to use alternative NLP packages before returning to scikit-learn to construct the models.

(Link : <https://scikit-learn.org/stable/>)

## **3.3 Chapter Summary & Evaluation**

In summary, this chapter discusses the research method and model that was designed and used to carry out the project. The methodology that is the machine learning life cycle act as the basis of the progress of this project. Research methods such as data collection and text cleaning play essential roles in sentiment analysis as data analysis is critical and will affect the results. In addition, several NLP libraries can help predict the sentiment from the text data throughout the entire project.

# Chapter 4

## **System Design**

## 4 Research Design

### 4.1 Data Visualization Design

By placing data in a visual context, such as maps or graphs, data visualisation helps us understand what it means. This makes the data more natural to understand for the human mind, making it easier to see trends, patterns, and outliers in vast data sets (Anon, 2020). By conveying data in the most effective manner possible, data visualisation may assist. Data visualisation is essential in the business intelligence process because it takes raw data, models it, and delivers it to conclude. Machine learning algorithms are being developed by data scientists in advanced analytics to better assemble critical data into visualisations that are easier to grasp and analyse.

Data visualisation, in particular, uses visual data to present information in a universal, quick, and effective manner. This method can assist businesses in determining which areas require improvement, which factors influence customer satisfaction and dissatisfaction, and what to do with certain items (where should they go and who should they be sold to). Stakeholders, business owners, and decision-makers can better estimate sales volumes and future growth using visualised data.

There are multiple techniques for putting information together in such a way that the data may be displayed. For example, several graphs and tables may be used to build an easy-to-understand dashboard, depending on the data being modelled and its intended purpose. In addition, some visualisations are made manually, while others are created automatically. In any case, there are a variety of options to satisfy the visualisation requirements.

In the age of big data, extracting information from text remains a difficult but vital challenge. The sheer number of data to be evaluated might overwhelm information to be extracted, whether it's from consumer feedback, social media posts, or the news. Modern natural language processing (NLP) techniques can help with this. They can analyse prevailing moods regarding a topic or product (sentiment analysis), summarise/classify essential topics from texts (summarisation/classification), and, remarkably, even answer context-dependent questions (like Siri or Google Assistant). Thanks to their development, individuals and businesses now have access to consistent, sophisticated, and scalable text analysis tools.

In this project, the data visualisation for the COVID-19 vaccination tweets will be visualised to readable and informative graphical representation to show the sentiment progress, data analysis, and sentiment result. Furthermore, a visualisation like charts, graphs, tables, and graphical statistics will be constructed using multiple visualisation tools such as Plotly, Bokeh, WordClouds etc.

### 4.1.1 Types of visualisation for sentiment analysis

#### Bar Chart

A bar chart shows categorical data as rectangular bars with lengths and heights that match each data point's values. Vertical or horizontal bars can be used to create bar charts. Bar charts help compare single or multiple categories of data. The bars can be combined to form a grouped bar chart when comparing more than one data category. Volume is used in bar charts to show changes between each bar. As a result, bar charts should always begin at zero. When bar charts do not start at zero, users are more likely to misinterpret the difference between data values.

For sentiment analysis, a bar chart will interpret the amount of positive or negative sentiment in the data. It can also illustrate the amounts of positive/negative words, frequency of words, amounts of tweets, etc. Figure 4.1.1-1 below shows an example of a bar chart that shows the number of sentiments.



Figure 4.1.1-1 : Bar chart showing the number of sentiments

#### Pie Chart

A pie chart depicts the relationship between parts and wholes in your data. Consider a pie, where each slice represents one component and the sum of all slices equals the entire. Humans draw to circles as the global emblem of oneness, perfection, and infinity, as Manuel Lima explains in *The Book of Circles*. We prefer things to be complete, and we want to know how the many parts go together. A pie chart provides the comfort of a part-to-whole relationship.

For sentiment analysis and NLP, pie chart can also be use to interpret the amount of positive or negative or neutral sentiment in the data. In addition, it can also be used to illustrate the amounts of positive/negative words, etc. Figure 4.1.1-2 below shows an example of a pie chart that shows the percentage of different sentiments in the data.

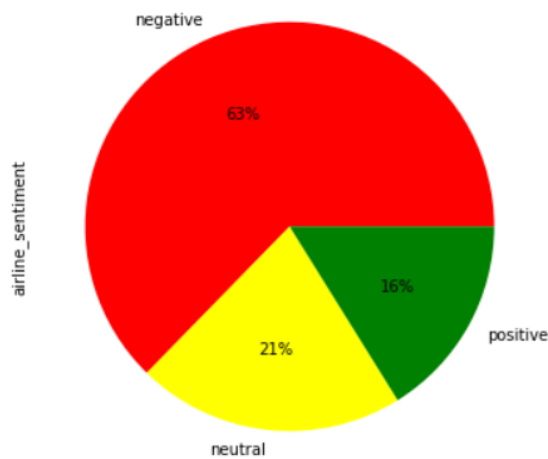


Figure 4.1.1-2 : Pie chart showing the percentage of different sentiments in the data.

## Word Clouds

The more often a specific word appears in a textual data source, the bigger and bolder it appears in the word cloud. A word cloud is a grouping of words that are displayed in various sizes. The larger and bolder the term, the more frequently it appears in a document and the more essential it is. These are ideal techniques to extract the most relevant sections of textual material, from blog posts to databases, and are also known as tag clouds or text clouds. They can also assist business users in comparing and contrasting two separate pieces of text to identify phrasing similarities.

For this sentiment analysis project, word clouds show the frequency of words in the data. The more frequently it appears in a document, the more essential it is and shown bigger in the word clouds visualisation. Figure 4.1.1-3 below shows an example of word clouds used to show the frequency of words in the data.



Figure 4.1.1-3 : Word clouds that shows the frequency of words in the data.

### Confusion Matrix

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. The confusion matrix itself is relatively simple to understand, but the related terminology can be confusing.

After cleaning, pre-processing, and wrangling the data, we first feed it to an excellent model and, of course, obtain output in probabilities. The Confusion Matrix is a performance metric for machine learning classification problems with two or more classes as output. There are four different combinations of projected and actual values in this table. It's great for measuring things like recall, precision, specificity, accuracy, and, most crucially, AUC-ROC curves.

The confusion matrix will be created once the sentiment model by using Naïve Bayes was tested to measure the accuracy and usability of the particular model. Then, it is to be used in the model evaluation stage to evaluate the best model out of others. Figure 4.1.1-4 below shows a confusion matrix with 4 different combinations of predicted and actual values.

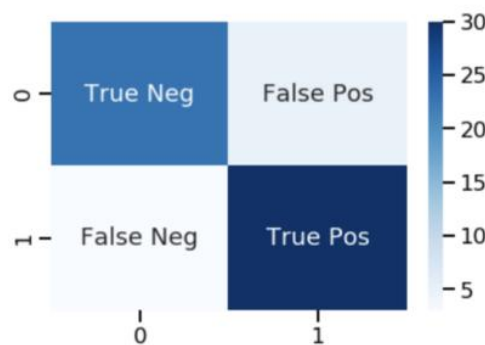


Figure 4.1.1-4 : Confusion matrix with 4 different combinations of predicted and actual values.

## 4.2 Model GUI Design

A GUI (Graphical User Interface) is a desktop application that allows you to communicate with computers. They carry out various jobs on computers, laptops, and other electronic devices. There are various graphical user interface (GUI) applications that we use regularly on our laptops and desktops.

Python provides several alternatives for creating a graphical user interface (Graphical User Interface). Tkinter is the most widely used GUI technique out of all the options. The fastest and most straightforward approach to construct GUI apps is with Python

and Tkinter. Tkinter is a Python tool for creating simple graphical user interfaces. In Python, it is the most often used module for GUI apps.

The final selected Naïve Bayes model will be deployed and have a GUI by using the Tkinter. The GUI can detect the sentiment of a sentence that is input by the user. Figure 4.2-1 below demonstrate an example of GUI as the sentiment detector.

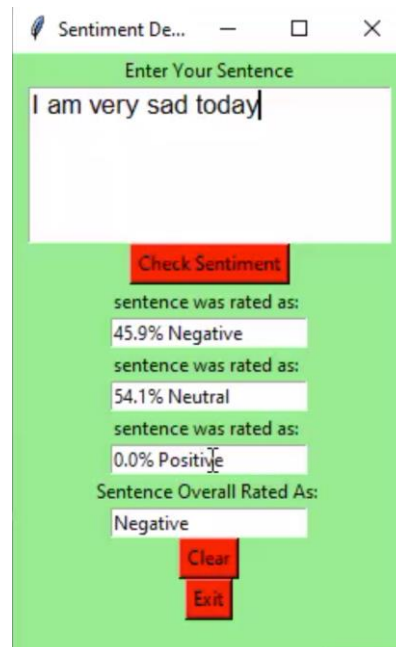


Figure 4.2-1 : GUI of a sentiment detector.

### 4.3 Chapter Summary and Evaluation

In summary, this chapter discusses the research design covering the data visualisation design and model GUI design. Several data visualisation designs are most proper for the use of sentiment analysis. Also, an example of the design of the GUI for the Naïve Bayes model has been shown.



## Chapter 5

# Results

## 5 Results

### 5.1 Test Plan

The test plan consists of the system testing and the experiment testing. The system testing is intended to determine whether the sentiment detector GUI system is capable of performing properly and meeting user needs. Functional testing, interface testing, performance testing, and end user testing will all be included in this testing phase. While the experiment testing in this project will be testing about the disparity of preprocessed data and non-preprocessed data when use for the data visualization to see the difference and the importance of data preprocessing.

#### 5.1.1 Objectives of Experiment Testing

Data pre-processing can be defined as a data mining approach that transforms raw data acquired from many sources into cleaner information that is more suitable for analyse and visualize. To put it another way, it's a preparatory phase that organises, sorts, and merges all of the available data.

If the phases of data pre-processing are skipped, the analysis' final output will be riddled with errors. This is especially true for more sensitive analysis that can be influenced by little errors, such as when it's applied in new fields where minor differences in raw data can lead to incorrect assumptions. Therefore, pre-processing the original dataset is crucial for this sentiment analysis research.

It is obvious why data pre-processing is so critical. Because errors, redundancies, missing values, and inconsistencies all jeopardise the set's integrity, you must address all of them for a more accurate result. Assume you're using a defective dataset to train a Machine Learning system to deal with your clients' purchases. The system is likely to generate biases and deviations, resulting in a bad user experience. As a result, before you use that data for your intended purpose, it must be as organised and "clean" as feasible.

Data pre-processing is a critical initial step for anyone working with large data sets. That's because it leads to better data sets, which are cleaner and easier to manage, which is essential for any company looking to extract useful information from the data it collects. So, the experiment testing of this project will be carried out to show the importance of data pre-processing by interpreting and displaying the analysis and results of data visualization by using the data that are not pre-processed.

#### 5.1.2 Objectives of System Testing

System testing comprises testing the entire system. All of the components are linked together to see if the system performs as planned. This is crucial for delivering a high-

quality output. System testing is done in a similar context to the production environment, so stakeholders may have a solid understanding of how the users will react. It aids in the reduction of post-deployment support and troubleshooting calls. This testing is critical and plays a key part in providing a high-quality product to the consumer.

The purpose of system testing is to reduce the risks associated with a system's behaviour in a certain environment. Testers do this by using an environment that is as near as feasible to the one in which a product will be installed when it is released. The objectives are a series of minor steps that help you reach your goal. There are various milestones in system testing that allow for the deployment of a flawlessly functioning system.

The following are the key objectives for a system testing:

- Lowering risks, because even bug-free components do not always work well as a system
- By performing a thorough investigation, as many faults and significant bugs as feasible are avoided.
- Verifying that the design, features, and performance of the product meet the requirements indicated in the product requirements.
- Before going on to the final stage – acceptability testing, which occurs just before users are given access to a product – the confidence in the system as a whole must be validated.

So, the system testing in this research project will be on the Sentiment Detector to test for its functional and also interface. It is to ensure that the final system is usable and meet the main objective that it serves.

## 5.2 Experiment Testing

Program Name :		Experiment Testing			
Test Date : 10/11/2021			Tester : Tan Yi Hong		
No.	Objective/Test Cases	Test Data	Expected Results	Actual Results	Remarks/Comments
1.	Importance of Data Preprocessing	Original dataset without preprocessed	The information and insights given by the Data Visualization on NLP does not contains meaning.	The word clouds and the graphs on NLP has fault insights.	Refer to Figure 5.2-1 until Figure 5.2-8.

Table 5.2-1 : Test plan for Experiment Testing

Figure 5.2-1 : Actual result of the word clouds for ‘Very Positive’ sentiment.

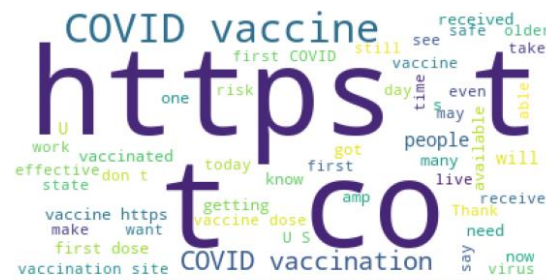


Figure 5.2-2 : Actual result of the word clouds for ‘Positive’ sentiment.

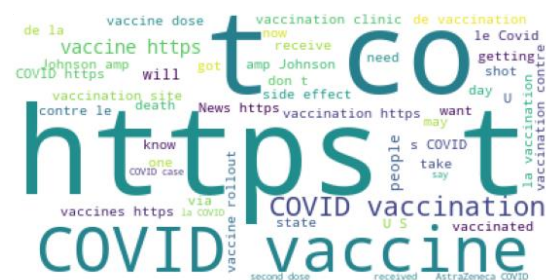


Figure 5.2-3 : Actual result of the word clouds for ‘Neutral’ sentiment.

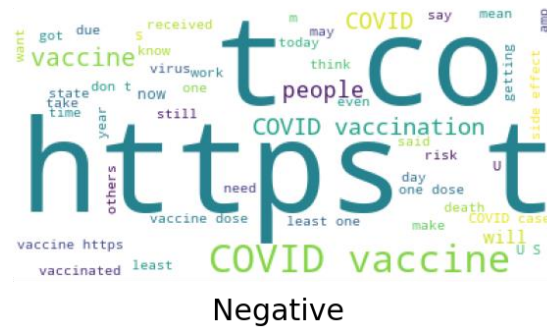


Figure 5.2-4 : Actual result of the word clouds for ‘Negative’ sentiment.

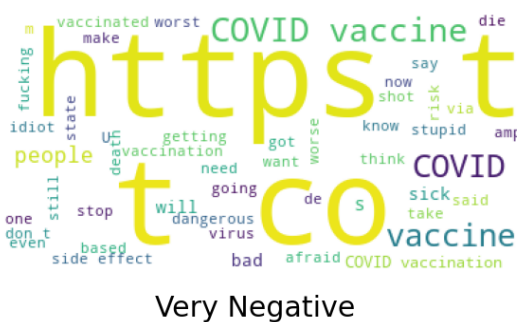


Figure 5.2-5 : Actual result of the word clouds for ‘Very Negative’ sentiment.

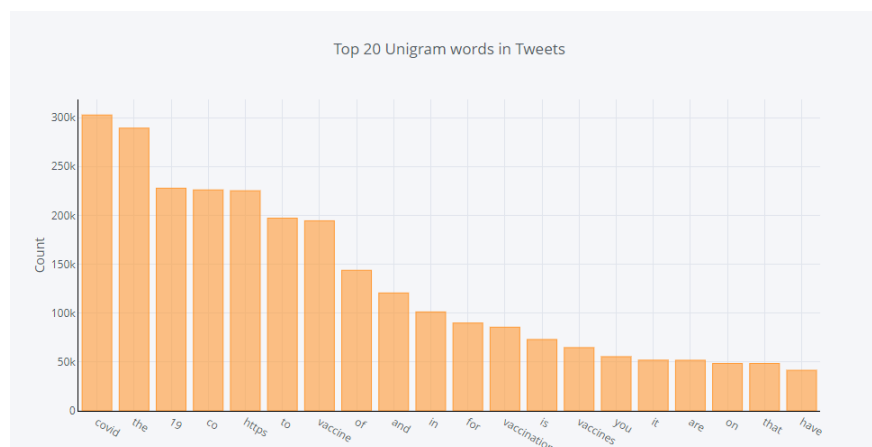


Figure 5.2-6 : Actual result of the bar chart for ‘Top 20 Unigrams words in Tweets’.

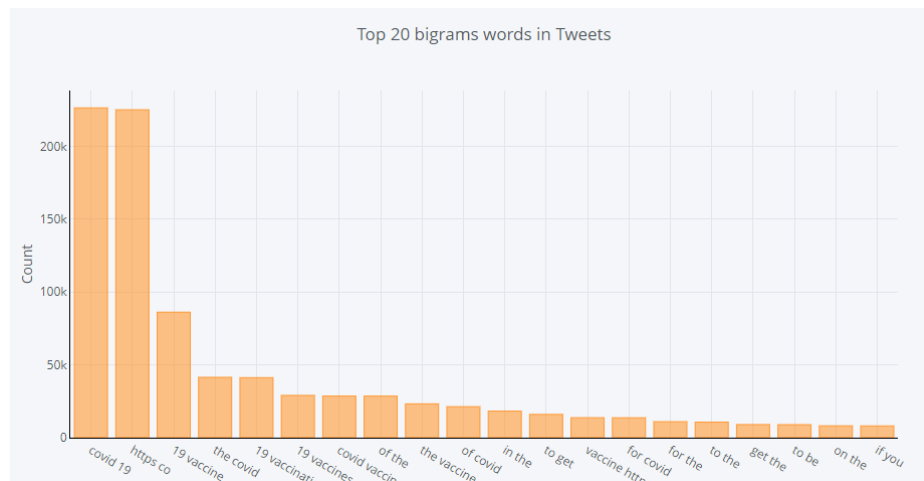


Figure 5.2-7 : Actual result of the bar chart for ‘Top 20 Bigrams words in Tweets’.

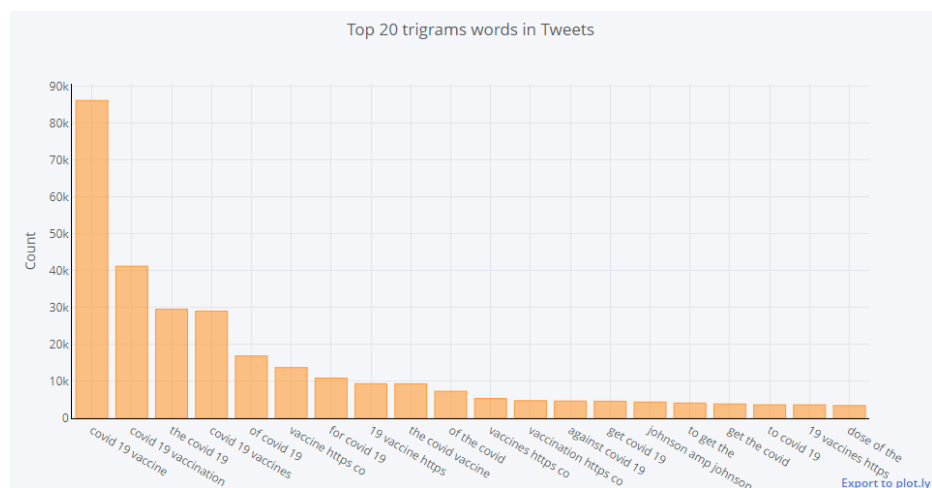


Figure 5.2-8 : Actual result of the bar chart for ‘Top 20 Trigrams words in Tweets’.

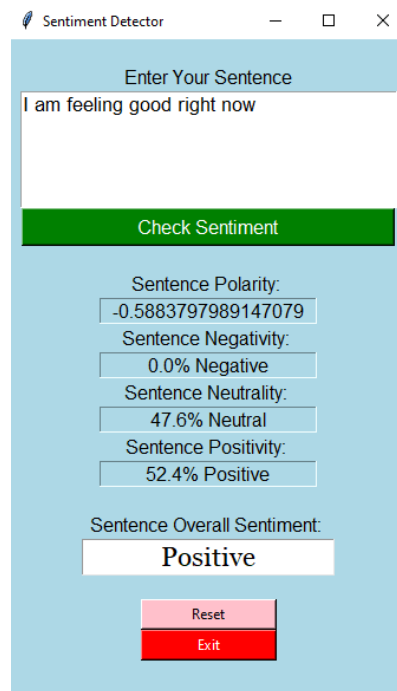
### 5.3 System Testing

Program Name :		System Testing			
Test Date : 10/11/2021			Tester : Tan Yi Hong		
No.	Objective/Test Cases	Test Data	Expected Results	Actual Results	Remarks/ Comments
1.	Positive Sentiment Detector	A positive sentence : “I am feeling good right now.”	The sentence overall sentiment will be rated as ‘Positive’.	The sentence overall sentiment is rated as ‘Positive’.	Refer to Figure 5.3-1

2.	Neutral Sentiment Detector	A neutral sentence : “The book has 40 pages”.	The sentence overall sentiment will be rated as ‘Neutral’.	The sentence overall sentiment is rated as ‘Neutral’.	Refer to Figure 5.3-2
3.	Negative Sentiment Detector	A negative sentence : “This glasses is not useful at all!”.	The sentence overall sentiment will be rated as ‘Negative’.	The sentence overall sentiment is rated as ‘Negative’.	Refer to Figure 5.3-3
4.	System Recoverability	User press the ‘Reset’ button.	The text input field will be cleared and results field will be empty.	The text input field is cleared and results field becomes empty.	Refer to Figure 5.3-4
5.	System Durability	User use it to test for 15 sentences	The system will be able to execute all sentences successful without having issues.	The system is able to execute all sentences successful without having issues.	
6.	System Performance	A random sentence as input.	The system will be able to detect the sentiment of the input sentence within 1 second.	The system detects the sentiment of the input sentence within 1 second.	

Table 5.3-1 : Test plan for System Testing

### Actual Results :

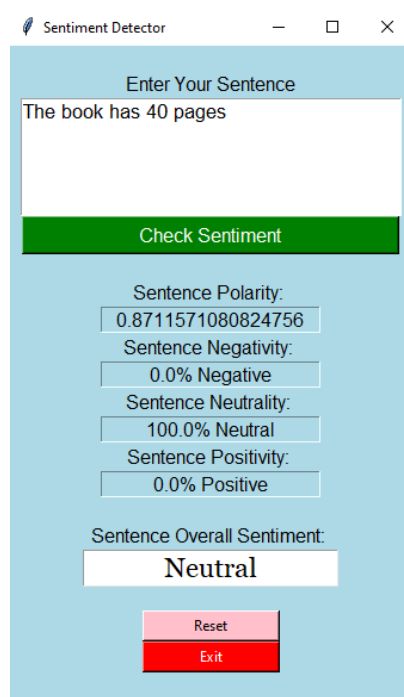


The screenshot shows a web application titled "Sentiment Detector". It has a text input field containing "I am feeling good right now". Below the input is a green "Check Sentiment" button. The results are displayed as follows:

Sentence Polarity:	-0.5883797989147079
Sentence Negativity:	0.0% Negative
Sentence Neutrality:	47.6% Neutral
Sentence Positivity:	52.4% Positive
Sentence Overall Sentiment:	Positive

At the bottom, there are two buttons: "Reset" (pink) and "Exit" (red).

Figure 5.3-1 : Actual result of Positive Sentiment Detector.



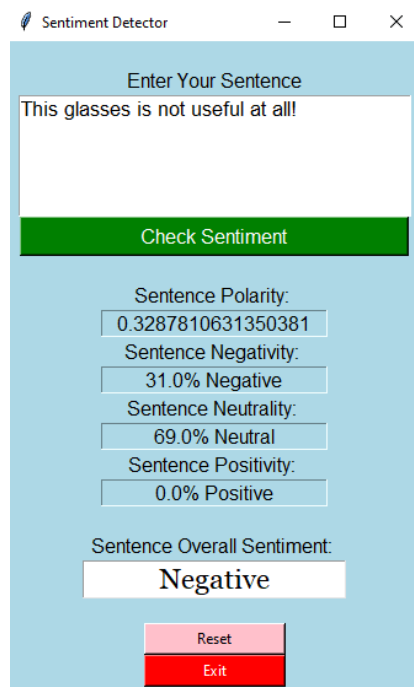
The screenshot shows the same "Sentiment Detector" application. The text input field now contains "The book has 40 pages". The results are displayed as follows:

Sentence Polarity:	0.8711571080824756
Sentence Negativity:	0.0% Negative
Sentence Neutrality:	100.0% Neutral
Sentence Positivity:	0.0% Positive
Sentence Overall Sentiment:	Neutral

The "Reset" and "Exit" buttons are still present at the bottom.

Figure 5.3-2 : Actual result of Neutral Sentiment Detector.



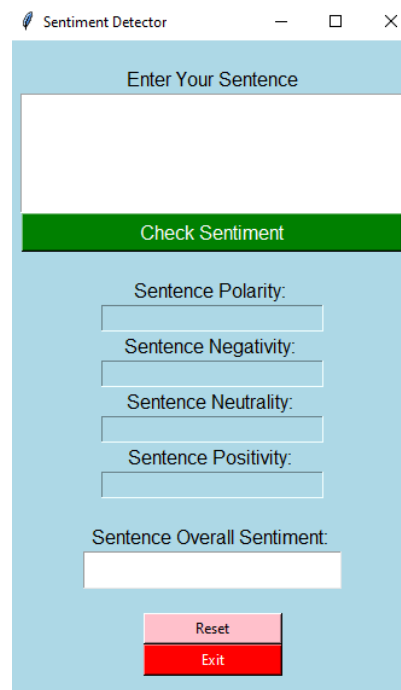


The screenshot shows a window titled "Sentiment Detector". Inside, there is a text input field containing the sentence "This glasses is not useful at all!". Below the input field is a green button labeled "Check Sentiment". Underneath the button, the following results are displayed:

- Sentence Polarity: 0.3287810631350381
- Sentence Negativity: 31.0% Negative
- Sentence Neutrality: 69.0% Neutral
- Sentence Positivity: 0.0% Positive
- Sentence Overall Sentiment: Negative

At the bottom of the results section, there are two buttons: a pink "Reset" button and a red "Exit" button.

Figure 5.3-3 : Actual result of Negative Sentiment Detector.



The screenshot shows the same "Sentiment Detector" window after the "Reset" button has been pressed. The text input field is now empty. The output fields for Sentence Polarity, Sentence Negativity, Sentence Neutrality, Sentence Positivity, and Sentence Overall Sentiment are also empty. The "Reset" and "Exit" buttons remain at the bottom.

Figure 5.3-4 : Actual result of System Recoverability after pressing 'Reset' button.

## 5.4 Chapter Summary and Evaluation

Summary of this chapter, the experiment plan is carried out in order to show the importance of data preprocessing for any Machine Learning or Natural Language Processing. Without data preprocessing, all of the data is unusable and carry 0 meaning for the data interpretation. While the system testing is crucial to test the entire system workflow to ensure that no bugs and error will happen when using the system.

## Chapter 6

# Discussions and Conclusion

## 6 Discussions and Conclusion

### 6.1 Summary

The COVID-19 pandemic has proven itself as one of the world's most dangerous threats, and it continues to be a source of concern. At the same time, we are in the midst of the world's largest vaccine campaign to combat the deadly virus. Unfortunately, while the vaccine has resurrected the fight against COVID-19, it has also triggered a wave of anti-vaccination rallies. So, it is important to gauge the public opinion towards the vaccination that has currently been given to almost all of the country in the world.

### 6.2 Achievements

Employing sentiment analysis on recent Twitter data by crawling Tweets data on vaccination on Twitter with the Twint API was beneficial in gauging public opinion on the COVID-19 vaccine. As a result, this study effort used the Nave Bayes approach conduct sentiment analysis and calculate the percentage of people's replies that are positive, neutral, or negative feelings about the COVID-19 vaccination for final categorization, as well as visualise the data and also a sentiment detector that will use VaderSentiment to calculate the sentiment compound of the sentence to predict the sentiment of the sentence accurately.

### 6.3 Contributions

In the current context, this sentiment analysis research study contributes as a vital source of better insights into public opinion about COVID-19 vaccination. This project used the Twint API to collect data from Twitter and show the results for analysis. The machine learning model will help to classify the types of sentiment, and the results will be visualised for everyone, notably government and scientists, to select their next action, such as holding a campaign, to encourage people to feel confident in the COVID-19 immunisation.

Data visualisation will be done throughout the project flow using various charts, graphs, or tables created with Plotly and Wordclouds to visualise the data process, analysis results, and overall insight of the sentiment analysis based on the COVID-19 vaccination in order to give a better image to the government, scientists, and general public.

### 6.4 Limitations and Future Improvements

There are a few limitations for this research project. Firstly, the data that we crawled does not carry many columns that can use for data visualisation. For example, the only usable columns of the tweets data are the date, tweets, replies\_count, retweets\_count, likes\_count, and etc. Usually for a sentiment analysis, the dataset should already contain

the sentiment label for each tweet to be interpret and visualize out. And here comes with another issue, which is the use of TextBlob to calculate the polarity of each sentence, and assign the sentiment according to the polarity. The polarity produced by the TextBlob for each sentence will sometimes not so accurate. For example, a sentence like 'I am not good today' should be a negative sentiment but instead, the TextBlob calculate the polarity of this sentence to be positive sentiment.

Therefore, the future improvements for these limitations would be, to crawl a dataset that contains more meaningful and interpretable columns so that I can explore more and visualize the insights of this dataset to be more interesting and informative. Next, the sentence sentiment will not be assigned base on only the polarity that calculated by the TextBlob, but will find more tools that can calculate and predict the sentiment more accurately.

## 6.5 Issues and Solutions

The main issue in this project is the prediction of the sentiment using the Naïve Bayes model. The Naïve Bayes model will calculate the polarity of the user input sentence, but it has a crucial problem is that the calculation of the polarity of the sentence is not so accurate. For example, a sentence 'She is not feeling well', the polarity will be less than 0 and will be predicted as negative statement. It is because the 'not' will not take into account when calculating the polarity. So, the final result of the sentiment detector will not be accurate.

To solve this issue, the solution is to not use the Naïve Bayes model to determine the sentiment result of the input sentence. By doing this, I had used the more accurate predictor which is VaderSentiment that calculate the sentiment compound of the input sentence, which means it can calculate the percentage of the sentence to which it was positive, neutral, or negative. This approach has successfully solved the problem mentioned above. For example, the same sentence, 'She is not feeling well', will be predicted as a negative sentence as the result.

## 7 References

A H Alamoodi, 2020, Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: A systematic review, viewed on 4 June 2021, <<https://pubmed.ncbi.nlm.nih.gov/33139966/>>.

Algorithmia, 2018, *Introduction to sentiment analysis: What is sentiment analysis?*, viewed on 29 May 2021, <<https://algorithmia.com/blog/introduction-sentiment-analysis>>.

Anon, 2020, *Why Data Visualization Is Important*, viewed on 4 July 2021, <<https://analytiks.co/importance-of-data-visualization/#:~:text=Data%20visualization%20gives%20us%20a,outliers%20within%20large%20data%20sets.>>.

Bahrawi Bahrawi, 2019, *SENTIMENT ANALYSIS USING RANDOM FOREST ALGORITHM ONLINE SOCIAL MEDIA BASED*, viewed on 4 June 2021, <[https://www.researchgate.net/publication/338548518\\_SENTIMENT\\_ANALYSIS\\_USING\\_RANDOM\\_FOREST\\_ALGORITHM\\_ONLINE\\_SOCIAL\\_MEDIA\\_BASED](https://www.researchgate.net/publication/338548518_SENTIMENT_ANALYSIS_USING_RANDOM_FOREST_ALGORITHM_ONLINE_SOCIAL_MEDIA_BASED)>.

BBC, 2020, *Covid-19 vaccine: First person receives Pfizer jab in UK*, viewed on 1 June 2021, <<https://www.bbc.com/news/uk-55227325>>.

C. Sari and Y. Ruldeviyani, 2020, *Sentiment Analysis of the Covid-19 Virus Infection in Indonesian Public Transportation on Twitter Data: A Case Study of Commuter Line Passengers*, viewed on 4 June 2021, <<https://ieeexplore.ieee.org/document/9255531>>.

Donato P M, 2021, *Covid-19 vaccines burnt as shelf-life complicates global rollout*, viewed on 1 June 2021, <<https://www.ft.com/content/8e1385cf-1569-4bf8-904e-fdc29367a758>>.

Heidi Ledford, 2021, *Six months of COVID vaccines: what 1.7 billion doses have taught scientists*, viewed on 1 June 2021, <<https://www.nature.com/articles/d41586-021-01505-x>>.

JKJAV, 2021, *National COVID-19 IMMUNISATION PROGRAMME*, viewed on 14 May 2021, <[https://www.vaksincovid.gov.my/pdf/National\\_COVID-19\\_Immunisation\\_Programme.pdf](https://www.vaksincovid.gov.my/pdf/National_COVID-19_Immunisation_Programme.pdf)>.

J Glob Health, 2020, *Key concerns in terms of the response to the COVID-19 pandemic*, viewed on 10 May 2021, <<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7297029/>>.

Koyel Chakraborty, 2020, *Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers—A study to show how popularity is affecting accuracy in social media*, viewed

on 4 June 2021, <  
<https://www.sciencedirect.com/science/article/abs/pii/S156849462030692X>>.

Muhammad Ali Fauzi, 2018, Random Forest Approach for Sentiment Analysis in Indonesian Language, viewed on 4 June 2021, <  
[https://www.researchgate.net/publication/327060733\\_Random\\_Forest\\_Approach\\_for\\_Sentiment\\_Analysis\\_in\\_Indonesian\\_Language](https://www.researchgate.net/publication/327060733_Random_Forest_Approach_for_Sentiment_Analysis_in_Indonesian_Language)>.

Nadirah H. Rodzi, 2021, *Anti-vaxxers turn vaccine advocates in Malaysia after brush with Covid-19*, viewed on 14 May 2021, <<https://www.straitstimes.com/asia/se-asia/anti-vaxxers-turn-vaccine-advocates-in-malaysia-after-brush-with-covid-19>>.

Nalini Chintalapudi, 2021, *Sentimental Analysis of COVID-19 Tweets Using Deep Learning Models*, viewed on 3 June 2021, <<https://www.mdpi.com/2036-7449/13/2/32>>.

OurWorldInData, 2021, *Coronavirus (COVID-19) Vaccinations*, viewed on 3 June 2021, <<https://ourworldindata.org/covid-vaccinations>>.

Rashida Nasrin Sucky, 2021, *Exploratory Data Analysis of Text data Including Visualisation and Sentiment Analysis*, viewed on 1 June 2021, <<https://towardsdatascience.com/exploratory-data-analysis-of-text-data-including-visualization-and-sentiment-analysis-e46dda3dd260>>.

Tania Jayatilaka, 2021, *The 5 Covid-19 Vaccines In Malaysia's National Vaccination Programme*, viewed on 10 May 2021, <<https://my.asiatatler.com/life/vaccines-in-malaysias-covid-vaccination-programme-pfizer-astrazeneca-sinovac-cansinobio-sputnik-v>>.

This page is intentionally left blank to indicate the back cover. Ensure that the back cover is black in color.