# Dashboard for Monitoring Depression Amongst Twitter Users Using Sentiment Analysis

by

Fatyn Nurraihanah binti Seman

17003222

Dissertation submitted in partial fulfilment of

the requirements for the

Bachelor of Information Technology (Hons)

SEPTEMBER 2021

Universiti Teknologi PETRONAS

32610 Bandar Seri Iskandar

Perak Darul Ridzuan

**CERTIFICATION OF APPROVAL**

# Dashboard for Monitoring Depression Amongst Twitter Users Using Sentiment Analysis

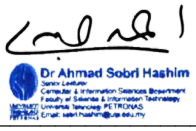
by

Fatyn Nurraihanah binti Seman

17003222

A project dissertation submitted to the
Information Technology Programme
Universiti Teknologi PETRONAS
in partial fulfilment of the requirement for the
BACHELOR OF INFORMATION TECHNOLOGY (Hons)

Approved by,




Dr Ahmad Sobri bin Hashim

UNIVERISTI TEKNOLOGI PETRONAS
TRONOH, PERAK
September 2021

## CERTIFICATION OF ORIGINALITY

This is to certify that I am responsible for the work submitted in this project, that the original work is my own except as specified in the references and acknowledgements, and that the original work contained herein have not been undertaken or done by unspecified sources or persons.

Fatyn Nurraihanah bt Seman

## ABSTRACT

This report will document and elaborate the idea of this project, Dashboard for Monitoring Depression Amongst Twitter Users Using Sentiment Analysis. By capturing Tweets from users on Twitter as the basis of research, sentiment analysis is performed by implementing machine learning to help identify the traits and polarity for depression categorised under mental health. The problem statement for this project being mental illness which has been one of the biggest health issues in Malaysia where sentiment analysis will be the tool to gather the proof from one of the many social media platforms, Twitter. The objective for this project is to use sentiment analysis technique for this case and prove the capability of the technique to identify and detect depression within Malaysian Twitter users. The method that had been use for this project is by implementing Machine Learning particularly to help assist the sentiment analysis technique in an agile flow. For the findings of this project, there will be some results in the form of percentage of people facing depression that will be placed into a dashboard format as well as some visualization on what the sentiment studies. To conclude, this project will support the findings and statement that had been released by UM Specialist Centre regards to the seriousness of mental health issue within Malaysian people.

## ACKNOWLEDGEMENT

At the very beginning, praise to Allah the Almighty for His blessings that I am capable to accomplish the completion of my Final Year Project 1 and 2. FYP is a compulsory subject that is required by Universiti Teknologi PETRONAS (UTP) students to complete the final year of their degree study.

First and foremost, I am indebted and thankful to numerous individuals for their kind advice, help, guidance and support which has encourage me to go through this challenging and enjoyable stage in my life learning for my degree. As well as being there for me to assist me throughout my journey and with my project.

Also, I would like to convey my sincere gratitude and deep respect to my supervisor, Dr Ahmad Sobri bin Hashim, for his assistance, insightful comments and information besides the guidance he had offered to me throughout this project. His kind gesture and supportive mentoring had helped me to push myself striving further and better for this project, without him I would have not been capable to complete the project within the time allocated.

I would also like to extend my thankfulness and appreciation to UTP and Computer and Information Sciences Department (CISD) committee for their excellent programme that had led me in such a great journey as their student. It is undoubtedly a fantastic opportunity for students to engage in the learning process by doing actual work and hands on learning objective which will help us prepare to face real-world difficulties and challenges in the future.

I would also want to express my heartfelt gratitude to my parents for their unwavering support and encouragement, which has inspired me to strive harder and become a better person. Thank you to my father, Hj. Seman bin Maimon for not giving up in me and trusted that I can achieve and become more than how much I thought of myself, it has helped me grow into the person I am today.

Last but not least, not to be forgotten, I would also like to express my gratitude to the people I call friends who have spent the last five years with me on this trip and sharing the ups and downs of student life. As well as to my fellow classmates BIT and BIS students of September 2018 for being such an amazing companion through this journey.

# Table of Contents

## LIST OF FIGURES

# CHAPTER 1

# INTRODUCTION

## 1.1 PROJECT BACKGROUND

Mental illness is a health condition that can be considered as a mental disorder when there are changes to the emotion, thinking or behaviour of a person. Mental illness affects mainly the brain of a person and is often correlated to stress. Mental illness is more than common, it can be affected by everyone; people of all age, gender, location, income, social status, race/ethnicity, religion, sexual orientation, background or culture (American Psychiatric Association, 2018). Mental illness is a serious matter because it can change a person completely and might as well ruin them.

Mental illness can make it hard for a person to cope with their daily life, work and routines as well as their relationship and other matters concerning an individual. Many people have at least experienced a mental health problem before as it is common and usually occurs from time to time. Mental illness would be a huge problem if it occurs for an ongoing duration of time and someone frequently experiences symptoms and episodes as it might cause more stress resulting in affecting a person's ability to act rationally. It is considered as a major issue that might lead to suicide which is stated to be one of the most common reason for the death of close to 800 000 people a year (World Health Organization, 2020).

Mental illness has no transparent or clear cause that can be identified as symptoms but there are several factors that could contribute to the growth of the mental health issues such as genes and family history, life experiences (traumatic events), biological factors, traumatic brain injury, exposure to viruses or toxic chemical before birth, alcohol or drugs, serious medical conditions and having few friends as well as feeling lonely and isolated. Research had been made and documented that mental illness can occur at any age, despite being known that three-fourth of the mental illness problem starts around the age of 24 (American Psychiatric Association, 2018).

There are many forms of mental illness from mild to severe which indicate the level of severity that has affected the individuals, despite whichever condition the person's state is mental health conditions are treatable. Individuals should seek for help to get better assistance and support whilst going through mental illness. At the moment, there is no certain cure for mental illness, although there are ways for a person to cope with the illness. In all countries of the world, the burden of mental diseases continues to rise, posing serious health risks as well as huge social, human rights, and economic repercussions. (World Health Organization, 2019).

As of today, there has been many attempts and steps that have been taken to determine if an individual has mental illness. Although most of mental illnesses are not "curable," they can be treated. Treatment for mental diseases varies widely depending on their diagnosis and the severity of their symptoms and the end result varies widely as well. Majority of people use medication, counselling or even both to help manage their mental illness. Some will need supplementary help with use of medication and some people require social support as well as guidance on how to manage their disease (MedlinePlus, 2021). There is no limitation or length of support to how much a person needs to help to get treated but there are many possibilities that can be carried out to help them find the best treatment for themselves (Mayo Clinic, 2019).

Technology evolves every day, it has become the centre of everything that had helped us in numerous ways from simplest task to the hardest task there can be. It has improved our daily lifestyles by providing us with more opportunities and tools to assist in our work and life which was beyond what had our imagination and thoughts before in the past. The evolvement of technology every day, changing and improving, has led to many chances for all industries to utilize this opportunity including the medical industry. The medical industry is now one of the most crucial sectors of the economy that is making use of latest technology as much as they could to assist in their daily tasks. Eventually, this has also led to the growth of many potential fields including the mental health and mental illness departments of the medical industry.

One of the technology advances that has been utilized by the medical field to assist them in mental health and mental illness is the implementation of machine learning to help them identify mental illness. Machine learning is a data analysis method that automates the development of analytical models. It is a type of artificial intelligence based on the notion that computers can learn from data, recognise patterns, and make judgments with little or no human input. Previous researchers have attempted to apply machine learning to the topic of mental health and have published their findings. Most research dataset were done based on Europe and Unites States societies, a wide spectrum of 75% on mental illness such as depression, anxiety, eating disorders, schizophrenia and many more was identified (Cho et al., 2019).

Sentiment analysis is a method that is done by a natural language processing technique used to determine whether data is positive, negative or neutral. Sentiment analysis can automatically analyse users or people words in social media and store it as a dataset based on the keywords. The contents from Twitter will be used as a data or material for the sentiment analysis project and be placed into the machine learning algorithm. Twitter is known as the social media platform where people can express themselves in many ways as well as use the like and retweet features. Machine learning will use some training data prior to being used as a training or study dataset to identify and state the polarity (positive, neutral or negative) and percentage of depression within the individual's tweet.

## 1.2 PROBLEM STATEMENT

According to UM Specialist Centre, mental illness is the second biggest health problem affecting Malaysian currently, one of the well-known mental illnesses is depression. We can get much insight on it through social media in which millennials are very active and expressive about with regards to their thoughts and feelings.

Regardless of the above, not much research has been done for it, as mental health is still stigmatized in our society.

### 1.2.1 Mental illness is second biggest health problem affecting Malaysians

According to the latest National Health and Morbidity Survey, one out of every three Malaysian adults aged 16 and up has a mental health problem (UM Specialist Centre, 2020). Unfortunately, there is still a dearth of knowledge about the condition, particularly in terms of typical mental disorders, their causes and consequences, and how to recognise the early indicators of significant mental difficulties (UM Specialist Centre, 2020).

### 1.2.2 Millennials expressiveness on social media

Nowadays, social media is the platform that is fully utilized by millennials. Social media has become the go-to application during their leisure time or when having something to post or express. Social media are interactive applications that allow people to create and share information, ideas, career interests and other kinds of expression through virtual communities and networks. Twitter is known as the social media platform where people normally express themselves more openly and expressively compared to physical social situations.

## 1.3 OBJECTIVES

The following are the objectives that will be attained by the end of the project:

- To propose sentiment analysis techniques to analyse sentiment express in tweets with regards to depression
- To validate the ability of the proposed technique in analysing the sentiment leading to depression
- To develop a dashboard that visualizes the sentiment analysis result into charts and listing

## 1.4 SCOPE OF STUDY

This project, like all others, requires a scope to describe the scope of the research and the parameters that will be used in the analysis. The scope of the project will be discussed further below:

- Malaysians

Malaysian citizens will be the subject of this project. Mental illness strike people all around the world, however the causes vary depending on a difference of the circumstances such as societal infrastructure, culture, geographic, ethics and so on. It is necessary to use Malaysian data and do studies on Malaysian characteristics to get a certain and accurate identification of mental illness among them.

- Twitter users

This project will be about sentiment analysis within the users of the social media platform Twitter. Twitter is a platform that is used to express themselves publicly to others whether it is by words, picture, videos or like and retweeting. Twitter users will be the target of this study to analyse the tweets that they posted whether its polarity will be associated to depression or not.

- Depression

As mentioned before, there are many types of mental illness such as depression, anxiety disorder, eating disorder and many more. For this project it would be used to highlight one of the most common mental illnesses which is depression. Research and study based will be made easier and simpler as it will be focused entirely on depression. Depression is described as the feeling of sadness, loss or anger within an individual that might resulted in the effect of their everyday performance.

These are some of the scope and perspectives that will be taken into account and used as a subject for this project. This is to help minimize and reduce the scope of study to get a better and more precise outcome.

# CHAPTER 2

# LITERITURE REVIEW

## 2.0 OVERVIEW

In this chapter, the project will be carried out based on research and review from literature review that could be found and related to the topic and scope of my project. This will help me to be able to analyse and evaluate previous research in order to give a literature review on the project topic area. There will be three parts that will be discussed here which is mental health, machine learning and sentiment analysis that will be the main focus of the study.

## 2.1 MENTAL HEALTH

### 2.1.1 DEFINITION

People's cognitive, behavioural, and emotional well-being are referred to as "mental health" and "behavioural health." It is all about how people think, feel, and behave. The term "mental health" is also used to describe the lack of a mental illness. Mental health, according to the WHO, is “more than merely the absence of mental diseases or disabilities”. It is described as a condition of well-being in which a person realizes his or her own potential, ability in managing typical life challenges, ability in working efficiently, and the ability to contribute to his or her community (World Health Organization & Felman, 2020).

### 2.1.2 MENTAL ILLNESS

Mental illness is a combination of abnormal thinking, perceptions, emotions, behaviour, and interpersonal connections characterises psychosis that affect mainly the brain of the individuals. It causes brain dysfunction as it affects the individual in terms of perception where their sensory system works in an unusual and/or unconventional ways that might play tricks on them. This also a brain dysfunction that affects the way they think either becoming abnormally fast or slow. Other than that, this brain dysfunction will also cause the mood of the individual to experience multiple and many mood changes from time to time easily. Lastly, the mental illness trigger brain

dysfunction that affect the behaviour and attitude in an odd and different way (Affairs, 1991).

Mental problems sometimes can start early in life and have a long-term recurrence as it has no age restrictions, from youngster to adults. They're common in every country where their presence has been studied. Mental diseases contribute significantly to the total disease burden due to their high prevalence, early onset, persistenceand disability (Hyman et al., 2011). In any given year, one out of every four persons suffers from mental health issues. The mental health issue ranges from common problems like depression and anxiety to more uncommon and rarer conditions like schizophrenia and bipolar disorder (Mind, 2017). Health services have yet to fully address the burden of mental diseases. Although there are known and effective treatments for mental disorders, between 76% and 85% of people in low and middle income countries receive no treatment for their mental disorder (Wang et al., 2007).

### 2.1.3 TYPE OF MENTAL ILLNESS

Mental illness is a term that classes all the type of illness that have affect towards an individual's behaviours, thinking mood and perception. The mental diseases come in a variety of forms. Some, such as specific phobia, are moderate and have little effect on daily life (abnormal fear). Other mental health issues are so serious that they may call for hospitalisation. Some of the types of mental illness are anxiety disorders, bipolar affective disorders, depression, eating disorders, schizophrenia and many more.

Anxiety disorder is a group of mental health illness that is caused by the rush of feelings of anxiety or fear by a situation or something. It can be based on phobias or angst of a particular environment or circumstances that will trigger the disorder. This could result in some symptoms such as shortness of breath, rapid heartbeat and dizziness. Anxiety disorder can be broken into three major disorders that are categorized as a part of it such as Generalized Anxiety Disorder (GAD), Panic

Disorder and Social Anxiety Disorder (SAD) (Morin, 2021). Bipolar disorder is when an individual feels the trigger episodes of mania, hypomania and depression. Their attitude and behaviour can shift dramatically depending on the episode they're going through which may or may not an experience of psychotic symptoms (BetterHealth, 2015). Although the actual cause is uncertain, a hereditary tendency has been proven for this illness.

Depression is characterised as a mood disorder where the person's feels sad, empty, lowered mood, loss of interest and excitement as well as loss of appetite. There are different type and symptoms of depression depending on the level of seriousness and severity as well as the symptoms of the depression. Depression is often related to anxiety disorder and if the illness is serious, it will increase the risk of negative impact such as suicidal thought or self-harm behaviours (BetterHealth, 2015). Eating disorder is also one of the mental illnesses which is defined as a long-term breakdown of eating patterns that result in poor physical and mental health. Eating disorder affects people from all ages, genders, social status and many more and it can result in serious psychological and physical consequence (BetterHealth, 2015). Anorexia, bulimia nervosa and other binge eating disorders are examples of eating disorders.

Schizophrenia is known as one of the rarer mental illnesses that is found diagnosed by people that are having mental illness. The schizophrenia is known as a psychotic syndrome characterized by disturbances in thinking and emotions, as well as a faulty perspective of reality which is consider as a serious brain illness. Hallucinations, delusions, mental illness, social withdrawal, lack of desire and impaired thinking and memory are some of the symptoms of schizophrenia. People with this disorder will have difficulties in keeping a job or having a career nor do they can take care of themselves welly. This had made people facing schizophrenia being a welly demotivated and a high risk to attempt suicide. Schizophrenia symptoms commonly appear between the ages of 16 and 30. Men are more likely than women to develop symptoms at a younger age. After the age of 45, most people do not develop schizophrenia (MedlinePlus, 2021b).

### 2.1.4 DEPRESSION

One of the most suffered mental illnesses amongst our civilization is depression. It can affect people of all ages, just like any other mental disorder. Depression is sometimes misunderstood or confused with sadness, although the two seem similar they are not the same. Sadness is a basic emotion that is commonly felt when a person is confronted with anything that makes them unpleasant. It is frequently a dominant emotion that causes a person to feel depressed or gloomy. Whereas depression is a different type of emotion as it is a persistent feeling of profound sadness when nothing appears to be wrong, but the person remains sad. A depressive episode can be categorized as mild, moderate or severe depending on the number of symptoms and intensity of it.

There are many symptoms of depression that can be a way of identifying a mental illness such as by their mood whether they feel sad or depressed. Next is that they might lose interest or excitement towards the things or activities that they enjoyed before. A change of appetite is also a symptom as the individual will either have a sudden loss of weight or gain it with an inconsistent diet. Besides that, individuals also tend to have trouble sleeping or have too much of it. Following then, having an excessive loss of energy or increase in fatigue is also a symptom of depression. Other than that, the person might develop an increase in the action of purposeless physical activities such as pacing and handwringing or a change in their pace on talking or movement. Also, the individuals might also have difficulties in thinking, concentrating and making decision due to their mental disorder. Lastly, the individuals also might have constant feelings like they are worthless or guilty which can lead to the thoughts of suicidal and death (American Psychiatric Association, 2020). Although the symptoms of depression that is experienced by men, women and children are all different.

According to the World Health Organization, around 264 million people worldwide suffer from depression particularly alarming for poor countries as 50.8 million people may be suffering from the illness (Azam et al., 2021). There are several possibilities to the cause of depression within these individuals as it spans from

biological to circumstantial. There is a higher chance of affecting individuals that have it in their family history. It is when family genes where in the family there is a history of depression or any other mood disorder. Next the depression might have been caused by an early childhood trauma or phobia that the individual was involved in which had cause them to react towards the event in fear and stressful situations. Besides that, the medical conditions of a person also can become a reason for the development of depression, conditions such as chronic illness, insomnia, chronic pain or attention-deficit hyperactivity disorder (ADHD). Other than that, the brain structure of an individual as well as misuse of drug or alcohol could led to depression (Higuera, 2020).

## 2.2 MACHINE LEARNING

### 2.2.1 DEFINITION

Machine learning is a subset of artificial intelligence (AI) that allows computers to learn and improve on their own without having to be explicitly programmed (Burns, 2021). Machine learning deals with the creation of computer programs that can access data and learn independently. The iterative feature of machine learning is crucial because models can evolve independently as they are exposed to fresh data. They use past computations to provide consistent, repeatable judgments and outcomes (Expert.ai Team, 2020).

### 2.2.2 TYPE OF MACHINE LEARNING ALGORITHM

Machine learning classification falls into three primary categories which are supervised machine learning, unsupervised machine learning and reinforcement machine learning. Each of these 3 types of Machine Learning algorithms functions and works in different ways. For supervised machine learning it is an algorithm that used dependent or result variable to be predicted from a set of predictors (independent variables). It creates a function that uses this set of variables to assign an input to the desired output. The model is trained until it achieves the appropriate degree of accuracy on the training data (Abdi, 2016). In this strategy, there is no target or result variable to forecast or estimate. It is widely used for segmenting clients into distinct groups for certain actions, as well as clustering populations into distinct groups (Abdi,

2016). Using this algorithm, the machine is taught to make particular decisions. This is how it works, the machine is placed in a situation where it must constantly teach itself through trial and error. In order to make suitable business judgments, this computer learns from its previous experiences and tries to capture the most relevant information (Abdi, 2016).

### 2.2.3 MACHINE LEARNING PIPELINE

The machine learning has a pipeline that serves as a roadmap or flowchart that elaborates or illustrates the flow of completing a machine learning project. Data retrieval, extraction, preparation, modelling, assessment and development are the six fundamental aspects of the machine learning pipeline. The machine learning pipeline is depicted in the diagram below:

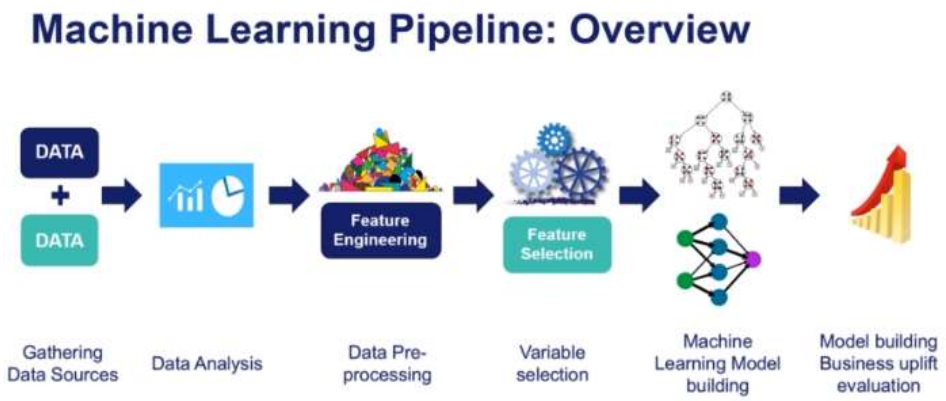



Figure 2.1: Machine Learning Pipeline

**Gathering Data Sources:**

The data source that will be used for this machine learning process on sentiment analysis will be from Twitter. Machine Learning is a crucial part of the picture. The machine learning method of Sentiment Analysis assists the system in understanding the sentiment of a given remark. The machine learning will be a system that will be used to extract the data from twitter on which sentiment analysis will be performed.

That data that has been pulled from Twitter is saved in a data frame. Twitter is one of the most popular data sources with a large volume of data. Without following any

restrictions, one is allowed to construct tweets in any format. This is one of the reasons why Twitter is more well-known than other blogging platforms (Yadav et al., 2021).

**Data Analysis:**

The data analysis phase is where the data will be examined to understand and answer questions such as:

- what variables are available?
- how are they related?
- what are the characteristics of those variables? (numerical or categorical?)
- missing values? outliers?

Because of the freedom and absence of restriction on how a tweet can be written, that Twitter permitted to their users the people are more likely to use abbreviations, misspell words, inflate evaluations and use emoticons, among other things. These formats make analysis more challenging, but there are still methods for investigating the tweets, such as feature extraction and mapping emoticons to their true meanings (Yadav et al., 2021).

**Data Pre-processing:**

For the data pre-processing phase, it is a step where the dataset undergoes processes such as filtering, transforming and identification to get the relevant data that is needed to be used. The objective of this phase is to let it undergo some cleaning and pre-processing procedures so that accurate data may be used to fit the machine learning model, which guides in the prediction of labels for unidentified data samples that have been cleaned and pre-processed (Yadav et al., 2021).

Some of the processes of data pre-processing are such as:

i. Evaluate the data quality.
ii. Identifying inconsistencies in values in order to determine what the data type of the features should be.

iii. Aggregate the features to give better performance

The keywords that will be identified for the sentiment studies is by selecting the dataset by words for depression such as “depress, lonely, sad, useless and others”.

**Variable selection:**

This step attempts to choose a subset of features from all features that are relevant to the performance of the ML model. This is significant because a large number of features might be present without feature selection, which is harmful to both model development and deployment. Feature selection is important as it highlights some of the following purpose(Shaikh, 2018):

- Reduces Overfitting: There will be less chance of making conclusions based on noise if lesser redundant data are presence in the dataset
- Improves Accuracy: This will help the model to have a better accuracy when there are lesser misleading data in the dataset
- Reduces Training Time: The complexity of algorithms is reduced when there are fewer data points, and algorithms train faster

**Machine Learning Model Building:**

Trying out different machine learning models and picking the best one for the job. However, it's worth emphasising that, in practise, this stage is only a small part of the overall pipeline and other processes are just as crucial, if not more so. The main idea is to use supervised learning techniques to detect patterns in known instances and apply them to new documents, allowing new documents to be classified automatically based on their sentiment orientation.

As a result, supervised learning aims to train and obtain an opinion classifier that includes some target opinion classes, such as positive vs negative (Luo et al., 2013). Some of the popular supervised learning that had been used for sentiment analysis are:

i. Naïve Bayes (NB): A straightforward probabilistic classifier based on Bayes' theorem and strong (naive) assumptions of independence.
ii. Maximum Entropy (ME): The conditional distribution of the class label is estimated using a probabilistic model.
iii. Support Vector Machines (SVM): Support vectors are computed to provide the optimal classification of points/instances into categories in a representation of an examples as a point in space.
iv. Logistic Regression Model (LR): A LR model predicts classes based on a set of continuous, discontinuous, or mixed variables.
v. Lexicon based approach: Calculates the sentiment orientation of an entire document or set of sentences from the semantic direction of the lexicon, uses adjectives and adverbs to find the semantic placement of the text.

## 2.3 SENTIMENT ANALYSIS

### 2.3.1 INTRODUCTION

Sentiment analysis is a method that uses Natural Language Processing (NLP) to extract, transform and analyse opinions from a text and classify them as positive, negative or neutral sentiment. It is used to help with decision-making in a variety of fields. Sentiment orientation classification refers to determining the opinion orientation of an opinionated text, determining whether the text's opinion orientation with polarity based on opinion words identified from specific text by emotional identification (Luo et al., 2013). Sentiment analysis solves these issues by systematically gathering and analysing online sentiments in real time from a big sample of customers that expressed their online feelings which are human convictions or emotions expressed on the internet (Rambocas & Gama, 2013).

Sentiment analysis has its roots in psychology, sociology, and anthropology, and stems from theories like affective stance and appraisal theory which emphasise the role of emotions in forming cognitions (Rambocas & Gama, 2013). Microblogging sites have grown in popularity as a source of a wide range of information. This is due to the nature of microblogs, which allow users to post real-time messages about their thoughts on a variety of topics, discuss current events, complain and express positive

sentiment for items they use every day (Srivastava et al., 2019). There are two types of SA techniques which are dictionary-based and machine learning-based. Dictionary-based techniques, often known as lexicon, are used to analyse polarity in text or words. Machine learning also includes the use of a training data set as well as classification features such as the order text of the post/data source (Prakash & Aloysius, 2020).

### 2.3.2 USES OF SENTIMENT ANALYSIS

Sentiment analysis derives together several study areas such as natural language processing, data mining and text mining, and is quickly gaining traction among businesses as they seek to integrate computational intelligence approaches into their operations and improve their goods and services (Farhadloo & Rolland, 2016). In general, sentiment analysis divides text expressions in source materials into two ways which are known as objective and subjective. Objective route is known as facts about entities, events and their attributes which is certain while the subjective route way is known as an opinion or perspective expressions of sentiments, attitudes, emotions, assessments or feelings regarding entities, events and their attributes (Luo et al., 2013).

One of the main goals of sentiment analysis is to recognise and classify choices in source text expressions such as to determine whether a particular document or a given sentence expresses opinions and whether those opinions are expressed as positive, negative or neutral (Luo et al., 2013). Despite the limitations of data structure and volume, marketers can learn about consumer thoughts and attitudes in real time with the help of sentiment analysis method (Rambocas & Gama, 2013). It has been stated that sentiment analysis has been used in the contexts of business and marketing, politics and public action for use of application such as E-commerce, voting applications and world events. The majority of the data for the study of sentiment analysis method was gathered from social media (Drus & Khalid, 2019).

Sentiment Analysis has been used and implemented for many applications and benefits the interest of businesses and organization in fields such as healthcare,

finance, hospitality, tourism, political, sports and others. For example, in healthcare, they use an application, SentiHealth-cancer for cancer context. SentiHealth-cancer is a technology that analyses the moods and emotions of cancer patients. The adverse drug approach is used to investigate the impact of emotions. For hospitality and tourism, sentiment analysis can also be used as a way to collect the reviews as tourists may now access a variety of sources of information and create their own comfortable environment and share their thoughts and abilities with a low cost as well as providing real time tool services (Prakash & Aloysius, 2020).

For political purpose, the sentiment analysis can be used to display the results and anticipate election outcomes using old-fashioned methodologies such as electoral surveys which helps to increase the election accuracy of politics (Prakash & Aloysius, 2020). Politicians now use it to assist in getting a better understanding the topics that matter to their people. For example, the strategy was used in the recent national election campaigns in the United Kingdom and the United States, and it is predicted to make a big breakthrough in the upcoming 2012 US elections (Rambocas & Gama, 2013).

Government entities in charge of homeland security can benefit from sentiment analysis. Agencies can obtain vital knowledge about new dangers by tracking surges in negative opinions about a specific political entity, authority or even country (Rambocas & Gama, 2013). Other than that, the sports industry also uses sentiment analysis to do analysis on user comments at Facebook or Twitter pages that had revealed men and women both express hard feelings like rage or terror, but there is a large difference in how they express soft feelings like joy or misery (Prakash & Aloysius, 2020).

# CHAPTER 3

# METHODOLOGY AND PROJECT WORK

## 3.0 OVERVIEW

The methodology that was implemented for this project will be detailed and in depth in this chapter, from the research phase until the project's completion. The Research Methodology, Project Development, Development Tools Used and the Project Activities conducted will all be covered in this chapter to present the flow of the project.

## 3.1 RESEARCH METHODOLOGY

Prior to the development of the project some preliminary investigations had been done to understand the basics of the project as well as the scope of the study that is related to my project. Other than that, research on the existing works had been done some of which are related to the project or its adjacent subjects to get further understanding on the project as well as topic areas related to it. This has helped me to identify the key elements of the project that will be crucial and help me to roughly outline the structure and flow of the project. My method of studies was by numbers or research and literature review.

The internet had been a wide network that had helped me to get number of resources and materials studies for the literature review and sites. Literature review is known as an existing article, research, findings or other resourceful sources that contain material that would provide the insights to my study. They are papers that are detailed and informative as they have extensive analysis, studies, assessment and synthesis that had been undertaken by past researchers or people. There are numbers of publication that had been posted and uploaded online which can easily be found in website such as Google scholar, ResearchGate, IEEE and many more.

Other than that, the materials online such as journal, web page, blogs and books are also some of the sources that had been used as a reference and source of studies.

The literature review and research had helped me to get a better understanding of the scope of my project as well as how to develop it. This is also based on the existing prototypes or models that are similar or related to the scope of my studies which had assisted me to visualize the end result as well as how to initiate and run the project.

## 3.2 DEVELOPMENT OF PROJECT

### 3.2.1 SYSTEMS DEVELOPMENT LIFE CYCLE (SDLC)

To accomplish a quality project and deliverables, a defined procedure and guideline must be implemented in a timely manner. SDLC phases have been used throughout the project completion process as a framework for identifying tasks performed at each stage of the software development process, which include planning, analysis, design, development, testing, implementation and maintenance phase. By implementing this life cycle, it helps me to develop my project in a manner that could enhance the software quality and the development process as a whole. This project development is broken down into two parts, the sentiment analysis and the dashboarding, the two parts are planned to be running concurrently in this agile methodology.


Figure 3.1: System Development Life Cycle

1. Planning

The planning phase is the phase that is used as the preliminary planning for the project details to be subjected prior to the project execution. The purpose of this phase is used to identify the problem and determine the solution of the project. This step is important as it will help me set up the project goals and description on what scope of studies as well as planning of the project. The planning phase helps me to prepare the idea of the project from start to end that can later on be used as a guideline for the project development as well as ensure that the project deliverables are met. To make sure the planning is followed through and moving at its pace, a timeline and Gantt chart are created that will be used to help me keep track of my project. During this planning phase, research on how sentiment analysis is done, what type of language and algorithm is used, how the data is extracted and which social media platform to extract such dataset as well as how the data is planned to be presented is drafted.

2. Analysis

The analysis phase is to look into the planning that had been decided and study more into depth of the scope of study for the project to have a better understanding of the project and the process needed to be done. To ensure the project flows well adequate information and details about the project is needed hence it is required that process such as gathering, analysing and identifying the subject and information needed for the project needs to be done. Based on that, research and study through the existing sentiment analysis research had been done as well as analysis on the dataset from Twitter. The dataset from Twitter will be based on polarity which had been identify by keywords that had been posted within the tweets. Studies on research and literature review is done to collect information regards sentiment analysis on depression using Twitter. Other than that, the dataset that had been identify will help to idealize and give points for what can be visualized in the dashboard.

3. Design

The design step is the most important phase before the dashboard is built as it requires sketching of the dashboard concept and designing of the data pipeline. This simplifies the development process later on because a fundamental understanding of how to

create and visualize the dashboard as well as its data flow has already been done. From the idea of the analysis that had been done, some rough sketches of what information that had been imagine can be visualize and presented into these sketches.

4. Development

All project execution takes place during the development phase. This is the most important phase in the SDLC because it is where the project development begins. In the development phase, all of the documentation from the preceding phases are transformed into the actual project. Running concurrently, when the data is pre-processed and featuring had been implemented the dashboard will be developed.

This phase will involve steps such as:

- Data extraction from existing dataset
- Training of Machine learning using the existing dataset
- Development of the sentiment analysis algorithm on depression
- Filtration and polarity studies done on the machine learning
- Extraction of live data from Twitter
- Calculation of the polarity and status of the tweets extracted
- Gathering of empirical data
- Visualization of gathered dataset

5. Testing

The project is then thoroughly tested throughout the testing phase and any flaws or errors are corrected. The project's output should match the predicted results as well as its objective in order to determine the project's success. As for my case the testing and development phase happens concurrently, project testing is frequently regarded as a component of the development process as well. Later, User Acceptance Testing (UAT) will be done with my supervisor to get the approval or feedback for further verification on the visualization prompted in the dashboard.

### 3.2.1.1 AGILE METHODOLOGY

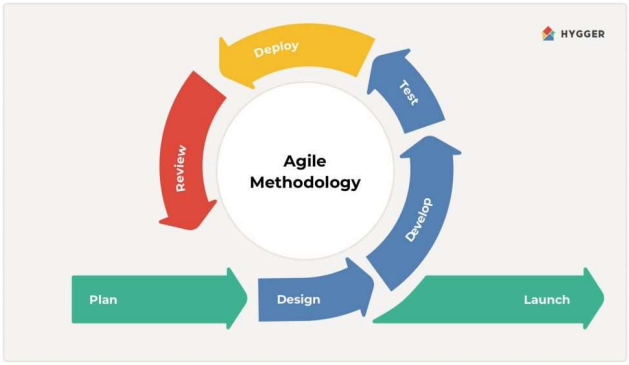



Figure 3.2: Agile Methodology

In the SDLC process, there are many models and approach that could be implemented, for my project specifically I would be implementing the agile methodology. Agile methodology is one of the iterative models in SDLC process that could be used for a project development. Agile methodology is a process known as the ideal strategy for many development teams, especially those aiming to create a continuous delivery environment, is to use a well-known development methodology. It emphasises iterative development, short development cycles, receiving feedback and adapting to new requirements. For this short timely project, I will be using this methodology to help me develop my project as I will be iterating from one phase to another to ensure that all steps are developing and timely allocated. Agile methodologies are divided to several steps which are plan, design, develop, test, deploy, review and then launch after the project is finalized.

### 3.2.2 SENTIMENT ANALYSIS PIPELINE

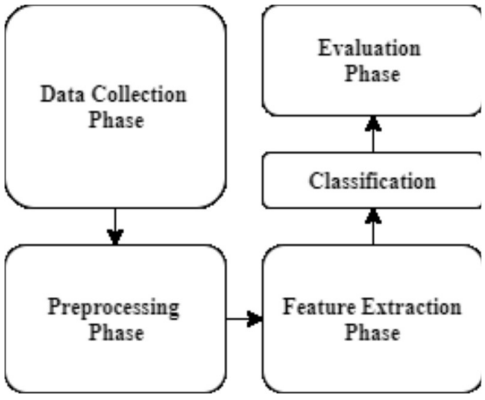



Figure 3.3: Sentiment Analysis Pipeline

1. Data Collection Phase

The data for this project is made up of a significant number of tweets gathered from the Kaggle repository and Twitter. The Twitter API is used to develop a Twitter application and obtain authorisation in order to gather tweets from Twitter. The information gathered comes in the form of both positive and negative tweets. The training dataset consists of both positive and negative tweets (Yadav et al., 2021).

2. Pre-processing Phase

The extracted raw tweets scraped from Twitter results in a noisy dataset. Tweets feature several unique characteristics, such as retweets, emoticons, client references and so on, that must be reasonably separated (Yadav et al., 2021). Some features such as URL, user mentions, emoticon and retweets or like are among the features that are available to be used by twitter users. To normalise the dataset and reduce its size, pre-processing processes were used such as reduce the size of the tweet by converting it to lowercase letters, use two or more dots should be replaced with spaces, remove any spaces or quotes from the end of the tweet, substitute a single space for two or more spaces and many more (Yadav et al., 2021).

3. Feature extraction

Unigrams and bigrams are two types of highlights that can be extracted from the dataset. For the retrieved features, a frequency distribution is produced. The top N unigrams and bigrams are then selected to do the analysis.

4. Classification

Selection of what algorithm that could be used and is most suited to the case study of the sentiment analysis, It will be involving the type of supervised machine learning, algorithm such as Naives Bayes, Support vector machine, Lexicon based approach and Random Forest.

### 3.3 DEVELOPMENT TOOLS

Following is some of the tools that is used to help develop the project:

i. Tools

Visual Studio Code


Figure 3.4: Visual Studio Code

Visual Studio Code is a software that combines the simplicity of a source code editor with advanced developer tooling such as IntelliSense code completion and debugging. Visual Studio Code is a cross-platform development environment that binds web, native and language-specific technology all in one platform. This tool will be use as the software and is a place where all the coding for the project will be done.

Anvil


Figure 3.5: Anvil

Anvil is a platform for building and hosting full-stack web applications written entirely in Python. It is based on drag and drop the UI and write Python on the front and back ends so that everything works. It has made web development much easier and faster. Anvil is a tool in the Platform as a Service category of the tech stack.

ii. Library and packages

Python


Figure 3.6: Python programming language

Python is a programming language that will be use for the coding in this project. It is a language that was created with readability in mind while having some similarities to English with mathematical impact. Its syntax enables programmers to create algorithms in lesser lines than other programming languages.

Tweepy

Figure 3.7: Tweepy

Tweepy is an open-source Python programme that made it very easy to access the Twitter API with Python. Tweepy contains a set of classes and methods that represent Twitter's models and API endpoints and it handles numerous implementation details transparently, such as:

- Encoding and decoding of data
- Requests made using HTTP
- Pagination of the results
- Authentication using the OAuth protocol

- Rate ceilings
- Streams

Natural Language Toolkit (NLTK)

The Natural Language Toolkit is a Python-based collection of tools and programmes that can work with human language for symbolic and statistical natural language processing (NLP) package for English. The documentation or text is an unstructured data with human-readable text making up a large portion of the data you could be examining. This includes classification, tokenization, stemming, tagging, analysis, word processing libraries for semantic thinking, industry standard NLP library wrappers, lively discussion forums, WordNet and more than 50 companies and lexical resources have included the simple interface. This library will help me filtered the sentences and identify the wording in the tweet that had been posted.

Matplotlib.pyplot

Matplotlib is a cross-platform data visualization and graphical plot library for Python and its numerical extension NumPy. Therefore, it provides a viable open-source alternative to MATLAB. This is a cross-platform library for creating 2D plots from the data in an array. It provides an object-oriented API to help you embed charts in your application using the Python GUI toolkit.

## 3.4 PROJECT ACTIVITIES

### 3.4.1 TWITTER SENTIMENT ANALYSIS

**Twitter Access and Authentication**

To start the development of the project, I analyse and understand first on how sentiment analysis is conducted and could be accomplished. For sentiment analysis to be able to work, it is required that the process is achieved by identifying where to extract the source of sentiment. The sentiment must be from an open source commonly social media where people usually express themselves publicly of the opinions and perspective they have in certain topic. Sentiment analysis could be implemented from

many platforms such as Facebook, Twitter, blogs and many more. For my case, I would be extracting the source of sentiments from Twitter as related to my problem statements on mainly millennials expressiveness towards having depression in social media. It has been proven that been proven statically that Twitter platform is a rather more popular platform for individuals to express themselves and 'confess' their feelings and thoughts. Statistics from Statista on "Distribution of Twitter users worldwide as of April 2021, by age group" as well shows that twitter have a wide range of millennials users which is an additional reason to choose Twitter as the source.

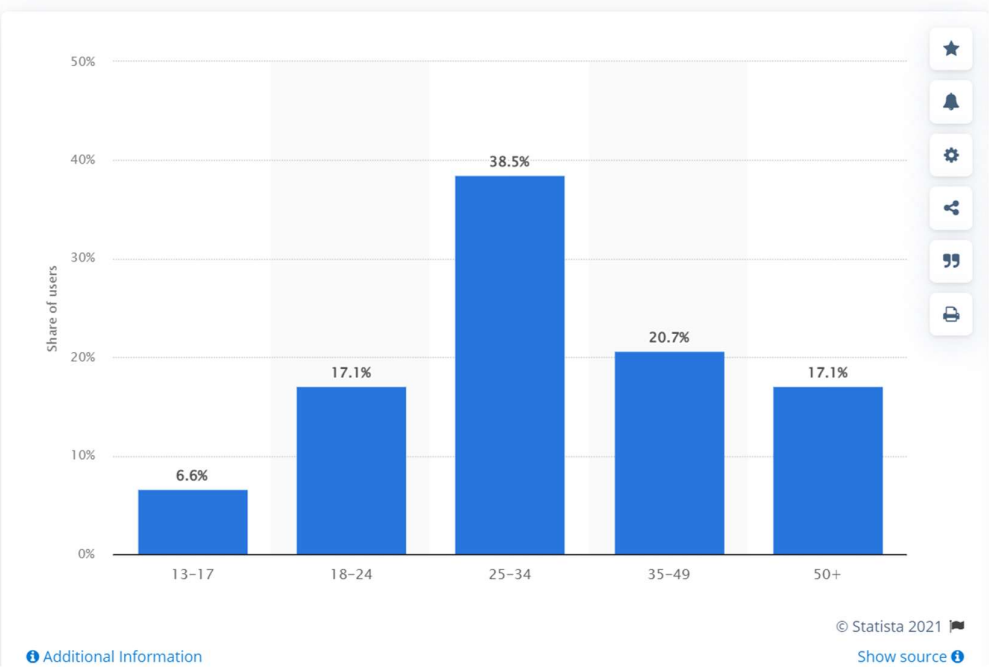



Figure 3.8: Twitter Users Age Group

Hence, to have access to Twitter users tweets it is required that I need to apply as Twitter Developer to be given permission to use the data and tweets for its API from the platform. The API is an application programming interface that acts as an intermediary between the developer's code and the Twitter database. All the data in the Twitter database requested by the developer goes through the API before being passed to the developer's code. To get a Twitter Developer account, however, developers must submit a request to Twitter that includes various details. Details such as:

- The core use case, intent, or business purpose for your use of the Twitter APIs.
- If you intend to analyse Tweets, Twitter users, or their content, share details about the analyses you plan to conduct, and the methods or techniques.
- If your use involves Tweeting, Retweeting, or liking content, share how you'll interact with Twitter accounts, or their content.
- If you'll display Twitter content off of Twitter, explain how, and where, Tweets and Twitter content will be displayed with your product or service, including whether Tweets and Twitter content will be displayed at row level, or aggregated

The application for the Twitter Developer account is done thoroughly and taken in a very detailed perspective to ensure that their user's privacy and content is protected without any violation for the requester. After case is taken into consideration and understanding, upon approval you will later on be given your very own Twitter API credentials that will permit you to have access to Twitter data.

**Sentiment Analysis Programming**

With the access permitted to retrieved data from Twitter, other steps are required to be determined to identify the best and most suitable way to start the project development. Sentiment Analysis is a machine learning process that analyse the polarity of text based on positive, negative and even neutral sentiments. It is known as a textual context mining that identifies and extracts subjective information in source material and helps businesses understand the social sentiment of a brand, product, or service while monitoring online conversations. There is various programming language that could be used as the language for the programming of this project. In the industry use, language like Java, C++, Python, R and many more had been utilized as the main languages to process sentiment analysis. For my choice, I would use Python as my programming language as it is known to be an easy language to be learnt as we well many build in libraries that would be helpful for my project development. As such libraries such as matplotlib, nltk, pandas, sklearn, textblob, numpy and many more had been used to assist me in the development of the project.

```
# Import libraries
import tweepy
from tweepy import OAuthHandler
from textblob import TextBlob
from wordcloud import WordCloud
import pandas as pd
import numpy as np
import re
import io
import csv
import matplotlib.pyplot as plt
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
import time
from nltk.sentiment import SentimentIntensityAnalyzer
import operator
from wordcloud import WordCloud, STOPWORDS, ImageColorGenerator
import copy
import string
from sklearn.feature_extraction.text import CountVectorizer
```

Figure 3.9: Imported libraries

Addition to that, python language has Tweepy which will assist in the process of retrieving and pulling tweets from Twitter much easier. Tweepy is well suited for collecting larger metadata, providing flexibility and potential scalability even when using the official API which will assist in the flow of the project as tweets are the main source of the project. As shown in the figure below of how Tweepy have been utilized, it has shortened the requirement and steps that is needed for me to code to extract tweets from Twitter.

```
16  # Get the twitter API credentials
17  consumerKey = 'insert your consumerKey'
18  consumerSecret = 'insert your consumerSecret'
19  accessToken = 'insert your accessToken'
20  accessTokenSecret = 'insert your accessTokenSecret'
21
22  # Create the authentication object
23  authenticate = tweepy.OAuthHandler(consumerKey, consumerSecret)
24
25  # Set the access token & secret
26  authenticate.set_access_token(accessToken, accessTokenSecret)
27
28  # Create the API object while passing in the auth info
29  api = tweepy.API(authenticate, wait_on_rate_limit = True)
30  places = api.geo_search(query="Malaysia", granularity="country")
31  place_id = places[0].id
32
33  searchcondition = ("place:%s" % place_id)
34  tweets = tweepy.Cursor(api.search , q=searchcondition, lang="en")
```

Figure 3.10: Tweepy utilization

Using Tweepy, I have implemented the *api.geo_search* function which will help me to filter all the tweets to only the geographical location of where the tweets had been posted Associated back to my scope of study which is to be based on Malaysian people sentiments on depression. The location has been specified to Malaysia so that the tweets that will be pulled are only from Malaysia and place into the query as the search condition for my tweet.

```
48  def get_tweets(query, count = 400):
49
50      # empty list to store parsed tweets
51      tweets = []
52      target = open('result.csv', 'w', encoding='utf-8')
53
54
55      with open('result.csv', 'w', newline= '', encoding='utf-8') as csvfile:
56          csv_writer = csv.DictWriter(
57          f=csvfile,
58         fieldnames=["Tweet"]
59      )
60          csv_writer.writeheader()
61
62      # call twitter api to fetch tweets
63      q=str(query)
64      a=str(q+" depressed")
65      b=str(q+" alone")
66      c=str(q+" tired")
67      d=str(q+" depression")
68      fetched_tweets = api.search(a, count = count)+ api.search(b, count = count)+ api.search(c, count = count)+ api.search(d, count = count)
69      # parsing tweets one by one
70
71      print(len(fetched_tweets))
```

Figure 3.11: Tweet Extraction

After setting up the condition that the tweet should only be from Malaysia, the tweets then should be extracted based on the keywords that will specify it more towards the

topic of the project which is Depression. Some common keywords of depression are as following:

| S.no | Word |
|---|---|
| 1 | Depression |
| 2 | Anxiety |
| 3 | Distressed |
| 4 | Demotivated |
| 5 | Insomnia |
| 6 | Lonely |
| 7 | Empty |
| 8 | Exhausted |
| 9 | Worried |
| 10 | Overwhelmed |
| 11 | Tired |
| 12 | Sad |
| 13 | Discouraged |
| 14 | Cry |
| 15 | Nervous |

Figure 3.12: Depression Keywords

Based from figure 9, the following are identified as the keywords for Depression studies for sentiment analysis which was referred from another article (Azam et al., 2021). I have taken 4 keywords to be included for the project query which are “depressed”, “alone”, “tired” and “depression”. These keywords will be included as the conditions to which the APIs will help me to extract based on the condition required from the tweets. The tweets will be filtered to the based-on tweets from Malaysia and which any that have any of the keywords placed in the query. For this project, I will extract 400 tweets that meets the requirement to be used as my reference to measure the sentiment of depression. The tweets that met the requirements will later then be stored in the ‘result.csv’ file.

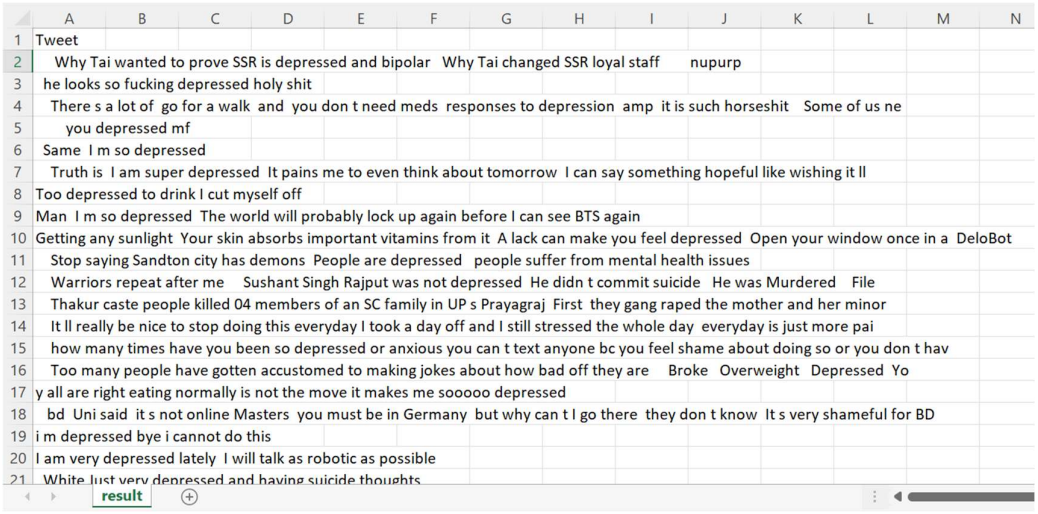
```
    for tweet in fetched_tweets:

        # empty dictionary to store required params of a tweet
        parsed_tweet = {}
        # saving text of tweet
        parsed_tweet['text'] = tweet.text
        if "http" not in tweet.text:

            line = re.sub("(@[A-Za-z0-9]+)|([^0-9A-Za-z \t])|(\w+:\/\/\S+)|(#)|(RT[\s]+)|(https?:\/\/\S+)", " ", tweet.text)

        target.write(line+"\n")

    return tweets

    # calling function to get tweets
tweets = get_tweets(query ="", count = 400)
```

Figure 3.13: Tweets Filtration

The tweet filtration process is conducted by removing any tweets that have hashtag, @user, link, retweets RT and such using regular expression. Following are the filtration that was used on each tweet that had met the conditions of the query before storing it into the csv file:

- '(@[A-Za-z0-9]+)'– removing username
- '([^0-9A-Za-z \t])' – removing unnecessary characters
- '(\w+:\/\/\S+)' – ensuring there is space and words
- '(#)' – removing hashtag
- '(RT[\s]+)' – removing retweets RT
- '(https?:\/\/\S+)' – removing https or links

After each tweet have been filtered, the tweets then later on is stored into the 'result.csv' file as meaningful tweets as it had removed all unnecessary data from the tweets that will affect the calculation of the sentiment score later on.



| | A |
|---|---|
| 1 | Tweet |
| 2 | Why Tai wanted to prove SSR is depressed and bipolar Why Tai changed SSR loyal staff nupurp |
| 3 | he looks so fucking depressed holy shit |
| 4 | There s a lot of go for a walk and you don t need meds responses to depression amp it is such horseshit Some of us ne |
| 5 | you depressed mf |
| 6 | Same I m so depressed |
| 7 | Truth is I am super depressed It pains me to even think about tomorrow I can say something hopeful like wishing it ll |
| 8 | Too depressed to drink I cut myself off |
| 9 | Man I m so depressed The world will probably lock up again before I can see BTS again |
| 10 | Getting any sunlight Your skin absorbs important vitamins from it A lack can make you feel depressed Open your window once in a DeloBot |
| 11 | Stop saying Sandton city has demons People are depressed people suffer from mental health issues |
| 12 | Warriors repeat after me Sushant Singh Rajput was not depressed He didn t commit suicide He was Murdered File |
| 13 | Thakur caste people killed 04 members of an SC family in UP s Prayagraj First they gang raped the mother and her minor |
| 14 | It ll really be nice to stop doing this everyday I took a day off and I still stressed the whole day everyday is just more pai |
| 15 | how many times have you been so depressed or anxious you can t text anyone bc you feel shame about doing so or you don t hav |
| 16 | Too many people have gotten accustomed to making jokes about how bad off they are Broke Overweight Depressed Yo |
| 17 | y all are right eating normally is not the move it makes me sooooo depressed |
| 18 | bd Uni said it s not online Masters you must be in Germany but why can t I go there they don t know It s very shameful for BD |
| 19 | i m depressed bye i cannot do this |
| 20 | I am very depressed lately I will talk as robotic as possible |
| 21 | White Just very depressed and having suicide thoughts |

result

Figure 3.14: Meaningful Tweets

The tweets are now more meaningful and useful to be measure for sentiment scoring as the tweets are now clear of any unnecessary and noisy data that might affect the scoring later on.

```
92  dataset = pd.read_csv('result.csv', header= None)
93  dataset.duplicated().sum()
94  dataset = dataset[~dataset.duplicated()]
95  dataset.duplicated()
96
97
98  dataset.columns = ['Tweets']
99  dataset.to_csv('dataset.csv')
```

Figure 3.15: Tweet Dataframe

Next step is to convert the tweets to a dataframe. The tweets are stored in result.csv file is then read and transferred into dataset as a dataframe. From the tweet that was extracted, there could be some similar tweets in the dataset as the process also takes retweets RT from a different individual as one tweet. Following is an example of a repetitive tweet that is stored in the 'result.csv' file:

```
47  H E  After working so hard  and spending nearly a year for our dream  now we are mentally depressed which is k
48   Piyali  Warriors repeat after me  Sushant Singh Rajput was not depresse
49  H E  After working so hard  and spending nearly a year for our dream  now we are mentally depressed which is k
50  H E  After working so hard  and spending nearly a year for our dream  now we are mentally depressed which is k
51 Pe0pels be like bro I m DEPRESSED  amp  SAD  like bro did you ever try to be HAPPY
52  Susan Walker didn t hv the qualification to diagnose depression yet she gave IVs to Lutyen journos declaring SSR as a d
53  Dr  Kersi Chawda in his statement to Bihar police denied Sushant being depressed  Why did bollywood try to peddle this
54 Bruh prioritizing my mental health over grades was a scam  I m still depressed  I m still anxious  now I m just dumb too
55 I am stressed and depressed now and it s because I m a virgin
56   Piyali  Warriors repeat after me  Sushant Singh Rajput was not depresse
57 I am  depressed
58 pray for space guys she s depressed
59  H E  After working so hard  and spending nearly a year for our dream  now we are mentally depressed which is k
```

Figure 3.16: Repetitive Tweets

As shown above on figure 13, tweets from row 47,49,50 and 59 have the same tweet. To remove redundancy, *dataset = dataset[~dataset.duplicated()]* is used so that the dataset will be cleared from any redundant data. After that *dataset.duplicated()* is used to check if there are still any duplication in the dataset. The dataset is then stored into another csv file which is 'dataset.csv' file. This will ensure that we would have only the data of original and unique tweets to be used for the sentiment scoring process.

```
104 from nltk.sentiment import SentimentIntensityAnalyzer
105 import operator
106 sia = SentimentIntensityAnalyzer()
107 dataset["sentiment_score"] = dataset['Tweets'].apply(lambda x: sia.polarity_scores(x)["compound"])
108 dataset["sentiment"] = np.select([dataset["sentiment_score"] < 0, dataset["sentiment_score"] == 0, dataset["sentiment_score"] > 0],
109                                  ['neg', 'neu', 'pos'])
```

Figure 3.17: Sentiment Scoring

Using the Sentiment Intensity Analyzer, I have categories the sentiments to be divided to three which are positive, neutral and negative. Positive sentiments will be the tweets that scores more than 0, neutral sentiments will be for the tweets that scores are equal to 0 while negative sentiments will be the tweets that scores less than 0.

```
111  from textblob import TextBlob
112
113  def getSubjectivity(text):
114      return TextBlob(text).sentiment.subjectivity
115
116  #Create a function to get the polarity
117  def getPolarity(text):
118      return TextBlob(text).sentiment.polarity
119
120  #Create two new columns 'Subjectivity' & 'Polarity'
121  dataset['Subjectivity'] =    dataset['Tweets'].apply(getSubjectivity)
122  dataset ['Polarity'] = dataset['Tweets'].apply(getPolarity)
```

Figure 3.18: Sentiment Subjectivity and Polarity

For additional information, sentiment subjectivity and polarity were also implemented to help measure the sentiments of the tweets better with the assistance of the sentiment scoring that was created.

```
126  #Pie chart
127  positive = dataset.loc[dataset['sentiment'] == 'pos'].count()[0]
128  neutral = dataset.loc[dataset['sentiment'] == 'neu'].count()[0]
129  negative = dataset.loc[dataset['sentiment'] == 'neg'].count()[0]
130
131  labels = ['Positive', 'Neutral', 'Negative']
132  colors = ['#add8e6', '#90ee90', '#ffcccb']
133
134  plt.pie([positive, neutral, negative], labels = labels, colors=colors, autopct='%.2f %%', textprops={'fontsize': 14})
135  plt.title('Sentiment of the Tweets')
136  plt.show()
```

Figure 3.19: Sentiments Pie chart

To help measure the percentage of each sentiment, a pie chart was illustrated to help visualize the portions.

- Positive – Non depression emotions
- Neutral – Neutral emotions, expression
- Negative – Depressive emotions

```
137
138  print("From total of ", dataset['Tweets'].count(), " tweets, there is ", positive,
139        " positive tweets, ", neutral, " neutral tweets and ", negative, " negative tweets.")
140
```

Figure 3.20: Sentiment Count

Following line of code was done to help identify how much tweets that are positive, neutral and negative. Below is the example of the counts:

```
From total of  228  tweets, there is  33  positive tweets,  11
neutral tweets and  184  negative tweets.
```

Figure 3.21: Example Sentiment Count

Then the sentiment tweets can then be ruled out so that we could identify what word was mostly mentioned in the tweets extracted. The frequency of each word will be count and could be visualize using wordcloud. Wordcloud are words that appear more frequently are given more emphasis in graphic representations. For this I have illustrated three kinds of wordcloud which are for overall, positive sentiment and negative sentiment words. Stopwords was used here so that they are unnecessary and meaningless words will not be included in the wordcloud.

```
142  from wordcloud import WordCloud, STOPWORDS, ImageColorGenerator
143
144  # Create stopword
145  stopwords = set(STOPWORDS)
146  # Generate a word cloud image for overall
147  allWords = ' '.join( [twts for twts in dataset['Tweets']])
148  wordcloud = WordCloud(width = 4000, height = 2000, random_state=1, background_color='black', colormap='Set2', collocations=False, stopwords
149  # Display the generated image
150  plt.imshow(wordcloud, interpolation='bilinear')
151  plt.axis("off")
152  plt.show()
```

Figure 3.22: Overall Wordcloud

```
154  # Negative wordcloud
155  neg_tweets = dataset.loc[dataset['sentiment'] == 'neg']
156  neg_string = []
157  for t in neg_tweets.Tweets:
158      neg_string.append(t)
159  neg_string = pd.Series(neg_string).str.cat(sep=' ')
160
161  wordcloud = WordCloud(width=1600, height=800,max_font_size=200, stopwords = set(STOPWORDS)).generate(neg_string)
162  plt.figure(figsize=(12,10))
163  plt.imshow(wordcloud, interpolation="bilinear")
164  plt.axis("off")
165  plt.show()
```

Figure 3.23: Negative Wordcloud

```
167  # Positive wordcloud
168  pos_tweets = dataset.loc[dataset['sentiment'] == 'pos']
169  pos_string = []
170  for t in pos_tweets.Tweets:
171      pos_string.append(t)
172  pos_string = pd.Series(pos_string).str.cat(sep=' ')
173  wordcloud = WordCloud(width=1600, height=800,max_font_size=200,colormap='magma', stopwords = set(STOPWORDS)).generate(pos_string)
174  plt.figure(figsize=(12,10))
175  plt.imshow(wordcloud, interpolation="bilinear")
176  plt.axis("off")
177  plt.show()
```

Figure 3.24: Positive Wordcloud

With this visualization it would also help to study what terms or words are usually used by users and what words mostly affects and plays a big factor to the sentiment scoring. After that, the dataset was then copied for further determination and frequency counts of words so that it could be studied. *Copy.deepcopy* was used so that the dataframe is copy out exactly without affecting the original dataframe.

```
202   import copy
203   #newdata = dataset[:]
204   newdata = copy.deepcopy(dataset)
```

Figure 3.25: Replicate Dataframe

To further the studies on word frequency and what relation connects the individuals to depression, the tweets will need to be cleanse off from all unnecessary words that might affect the counting process. Following are the cleaning process that was implemented in the project:

```
211   #Removing Punctuation
212   def remove_punct(text):
213       text = "".join([char for char in text if char not in string.punctuation])
214       text = re.sub('[0-9]+', '', text)
215       return text
216   newdata['punct'] = newdata['Tweets'].apply(lambda x: remove_punct(x))
217
218   #Appliyng tokenization
219   def tokenization(text):
220       text = re.split('\W+', text)
221       return text
222   newdata['tokenized'] = newdata['punct'].apply(lambda x: tokenization(x.lower()))
223
224   #Removing stopwords
225   stopword = nltk.corpus.stopwords.words('english')
226   def remove_stopwords(text):
227       text = [word for word in text if word not in stopword]
228       return text
229   newdata['nonstop'] = newdata['tokenized'].apply(lambda x: remove_stopwords(x))
230
231   #Appliyng Stemmer
232   snowball = nltk.SnowballStemmer(language='english')
233   def stemming(text):
234       text = [snowball.stem(word) for word in text]
235       return text
236   newdata['stemmed'] = newdata['nonstop'].apply(lambda x: stemming(x))
```

Figure 3.26: Cleaning Process

There are many filtrations that are required to be implemented so that the word frequency count will not be affected, some process such as removing punctuation, applying tokenization, removing stopwords and applying stemmer was applied.

- Removing punctuations: This code cleans up a string by removing any punctuation marks. Using the for loop, we will examine each character of the string. If the character is a punctuation mark, it is given an empty string.
- Applying tokenization: Tokenization in Python is the process of breaking down a huge body of text into smaller lines, words, or even inventing new terms for a language other than English. The nltk module includes a number of tokenization routines that can be used in programmes, as demonstrated below.
- Removing stopwords: Stopwords are English words that don't add much to a sentence's meaning. It can be safely ignored without compromising the meaning of the sentence, words like the, he, have are example of stopwords.
- Applying stemmer: A method to reduce the change in the ending of a word into its root form.

For the stemming process, snowball stemmer was used to help root out the words as snowball stemmer is proven to be slightly better comparing to the porter stemmer as there was some improvement was done to fix some of the issue present in the porter stemmer version.

```
#Cleaning Text
def clean_text(text):
    text_lc = "".join([word.lower() for word in text if word not in string.punctuation]) # remove puntuation
    text_rc = re.sub('[0-9]+', '', text_lc)
    tokens = re.split('\W+', text_rc)    # tokenization
    text = [snowball.stem(word) for word in tokens if word not in stopword]  # remove stopwords and stemming
    return text
newdata.head()
```

Figure 3.27: Cleaning Implementation

The functions were later on placed into one function which is *clean_text* to help clean the tweets from all unnecessary wordings.

```
from sklearn.feature_extraction.text import CountVectorizer

#Appliyng Countvectorizer
countVectorizer = CountVectorizer(analyzer=clean_text)
countVector = countVectorizer.fit_transform(newdata['Tweets'])
print('{} Number of reviews has {} words'.format(countVector.shape[0], countVector.shape[1]))
#print(countVectorizer.get_feature_names())

count_vect_df = pd.DataFrame(countVector.toarray(), columns=countVectorizer.get_feature_names())
count_vect_df.head()

# Most Used Words
count = pd.DataFrame(count_vect_df.sum())
count = count.reset_index()
countdf = count.sort_values(0,ascending=False).head(20).reset_index(drop=True)
countdf.columns = ['words', 'count']
print(countdf[1:14])
```

Figure 3.28: Word Frequency Implementation

The function was later then applied in the *countvectorizer* so where the words will be counted one by one from each tweet. This process which is shown in figure 25 will process all the words that are available in each tweet and count them. If the word is present more in other following tweets the counter will increase. In the end result *countdf* was called to print out the top 10 words. *Countdf[1:14]* was done so that keywords used for the query will not be accounted. Below is the example of the word counted for based on its frequentness in the tweets.

| | words | count |
|---|---|---|
| 1 | depress | 107 |
| 2 | tire | 53 |
| 3 | alon | 51 |
| 4 | peopl | 17 |
| 5 | get | 15 |
| 6 | make | 13 |
| 7 | im | 13 |
| 8 | like | 13 |
| 9 | go | 12 |
| 10 | one | 11 |
| 11 | feel | 11 |
| 12 | amp | 10 |
| 13 | realli | 9 |

Figure 3.29: Top 13 Words

```
# bar plot for the freq of words count
fig, ax = plt.subplots(figsize=(8, 8))

# Plot horizontal bar graph
countdf.sort_values(by='count').plot.barh(x='words', y='count', ax=ax, color="blue")
for p in ax.patches:
        ax.annotate(f'\n{p.get_height()}', (p.get_x() + 0.4, p.get_height()), ha='center', va='top', color='white', size=18)
for p in ax.containers:
        ax.bar_label(p, label_type='edge')
ax.set_title("Common Words Found in Tweets")
plt.show()
```

Figure 3.30: Bar plot of Word Frequency

Based on the list shown in figure 26, the *countdf* is later on used to draw a bar plot of the word frequency so that we could identify it better in comparison with other words.

```
#Function to ngram
def get_top_n_gram(corpus,ngram_range,n=None):
 vec = CountVectorizer(ngram_range=ngram_range,stop_words = 'english').fit(corpus)
 bag_of_words = vec.transform(corpus)
 sum_words = bag_of_words.sum(axis=0)
 words_freq = [(word, sum_words[0, idx]) for word, idx in vec.vocabulary_.items()]
 words_freq =sorted(words_freq, key = lambda x: x[1], reverse=True)
 return words_freq[:n]

#n2_bigram
n2_bigrams = get_top_n_gram(newdata['Tweets'],(2,2),20)
print(n2_bigrams)
```

Figure 3.31: Two Words Count

Another process that was done is counted out two words that goes together based on all the tweets that had been extracted. This code shown in figure 28, will address the two words that are together in a tweet frequency in the tweets. Following is the example of the end result of the two words count:

```
[('depressed didn', 3), ('im tired', 3), ('post concert', 3), ('concert depression', 3), ('just feel', 2
), ('like depressed', 2), ('didn know', 2), ('piyali warriors', 2), ('warriors repeat', 2), ('repeat sus
hant', 2), ('sushant singh', 2), ('singh rajput', 2), ('commit suicide', 2), ('sure depressed', 2), ('do
n wanna', 2), ('person okay', 2), ('okay powerful', 2), ('powerful person', 2), ('psychology says', 2),
('gt gt', 2)]
```

Figure 3.32: Example of Two Words Count

### 3.4.2 DASHBOARDING

Finishing with the sentiment analysis and steps required, the illustration and details are later on passed to web dashboard server. For my project, I utilized Anvil, Anvil is a Python-based platform that allows developers to build a full-stack web application. Anvil's platform supports drag-and-drop UI organisation, which saves developers a lot of time when it comes to writing the entire UI. Developers can deploy their app with a single click after they've finished building it. Anvil serves as a dashboard in this project, displaying the sentiment analysis results generated by the backend. Following are the steps that are conducted to visualize and show the details that had been made from the backend coding of this project:

1. Create an application for the dashboard

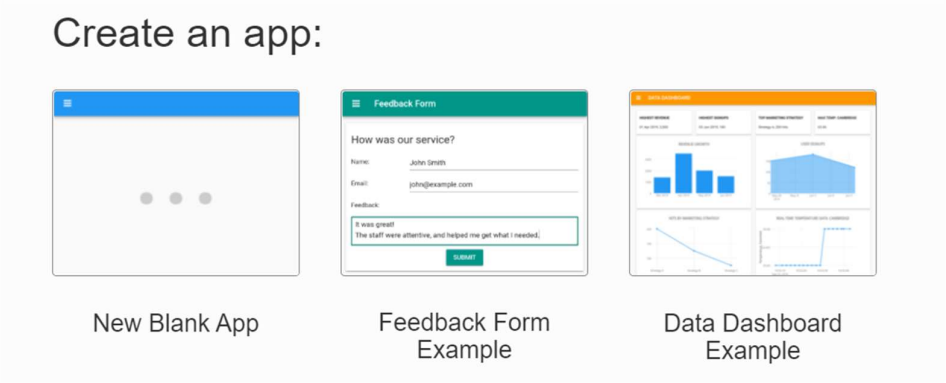



Figure 3.33: Create Dashboard

2. Designing the page

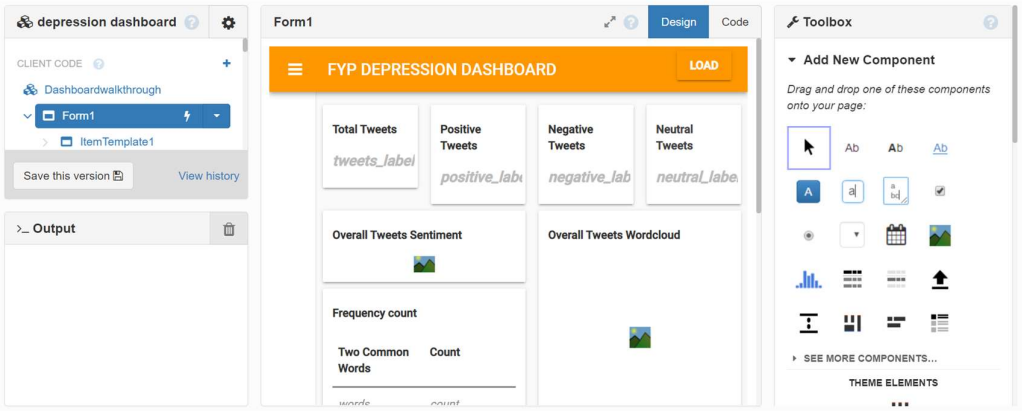



Figure 3.34: Design Dashboard

3. Add button and add on its events to call out the illustrations

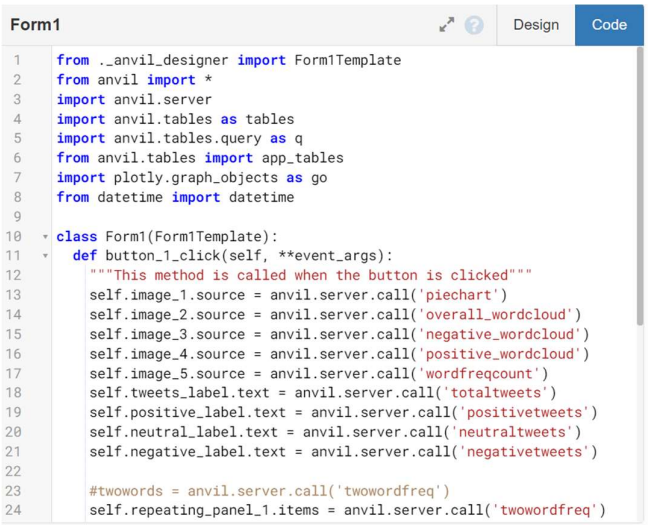



Figure 3.35: Button event

4. Retrieve the uplink to the dashboard server

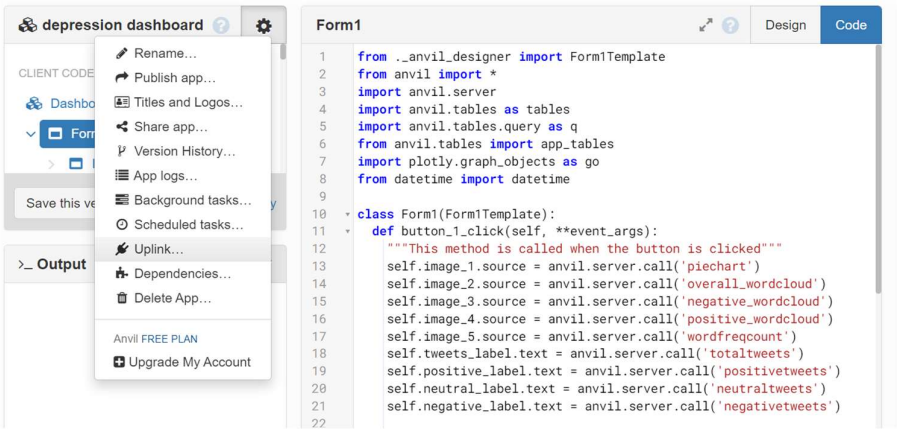



Figure 3.36: Establishing uplink connection

5. Connect the backend code script to the server

```
import anvil.server
import anvil.mpl_util

anvil.server.connect("uplink key insert here")
```

Figure 3.37: Connecting backend to dashboard

6. Create callable function that can be call and used from the button event

```
#Positive tweets
@anvil.server.callable
def positivetweets():
    positive = dataset.loc[dataset['sentiment'] == 'pos'].count()[0]
    print(positive)
    return positive

print(positive)
#Neutral tweets
@anvil.server.callable
def neutraltweets():
    neutral = dataset.loc[dataset['sentiment'] == 'neu'].count()[0]
    print(neutral)
    return neutral
print(neutral)
#Negative tweets
@anvil.server.callable
def negativetweets():
    negative = dataset.loc[dataset['sentiment'] == 'neg'].count()[0]
    print(negative)
    return negative
print(negative)
#Total tweets
@anvil.server.callable
def totaltweets():
    total = dataset['Tweets'].count()
    print(total)
    return total
```

Figure 3.38: Callable functions

Based on the figure 33 above, the code is being retrieved by applying *@anvil.server.callable* so that the server could recognise the function that is being called. Other than that, *anvil.server.wait_forever()* is also used so that Anvil can call the function endlessly because the backend is kept running.

7. Publish the dashboard

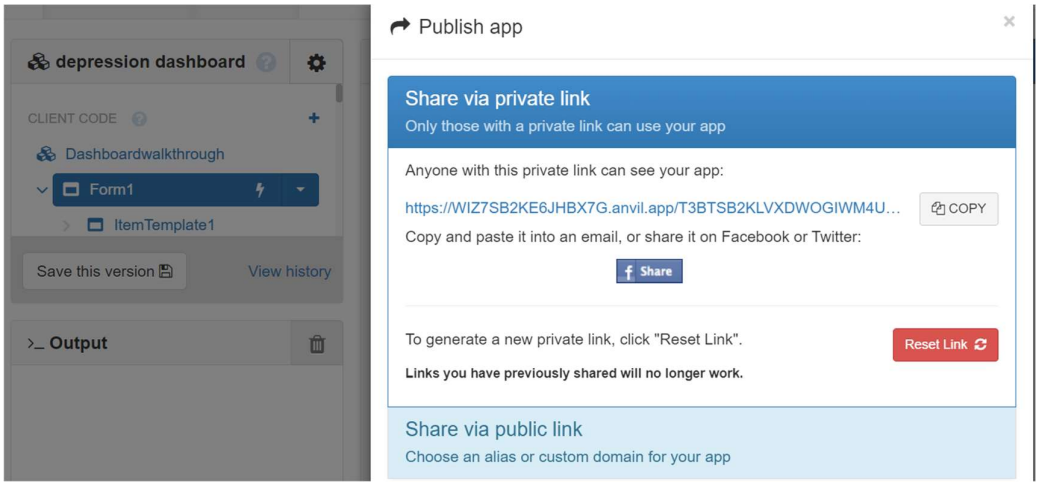



Figure 3.39: Dashboard Publishing

# CHAPTER 4

# RESULTS AND DISCUSSION

## 4.1 RESULT

Based on the project activities that was explained above there are several illustration and details that will be shown as an end result. Following are the things that will be shown as the result from this project:

i. Total number of tweets
ii. Total number of positive tweets
iii. Total number of neutral tweets
iv. Total number of negative tweets
v. List of word frequency
vi. List of two-word frequency
vii. Pie chart of tweets sentiment

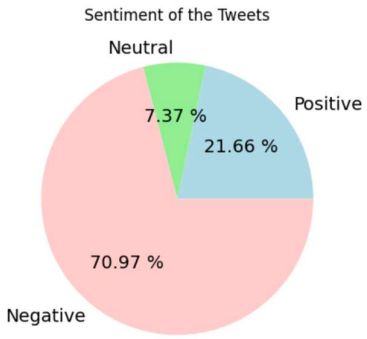



Figure 4.1: Example of Pie chart

viii. Overall Wordcloud

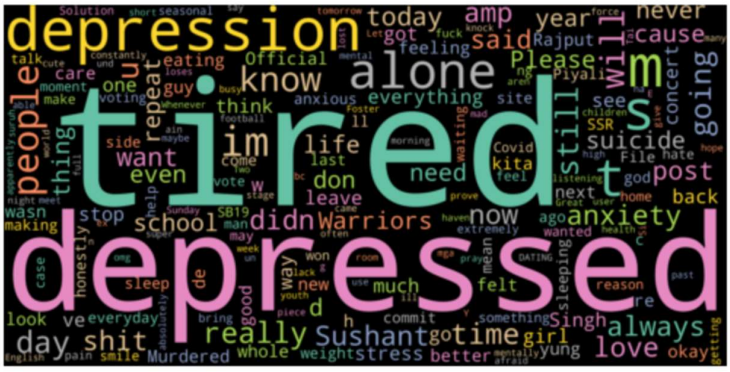


Figure 4.2: Example of Overall wordcloud

ix. Positive sentiment wordcloud


Figure 4.3: Example of Positive Sentiment Wordcloud

x. Negative sentiment wordcloud

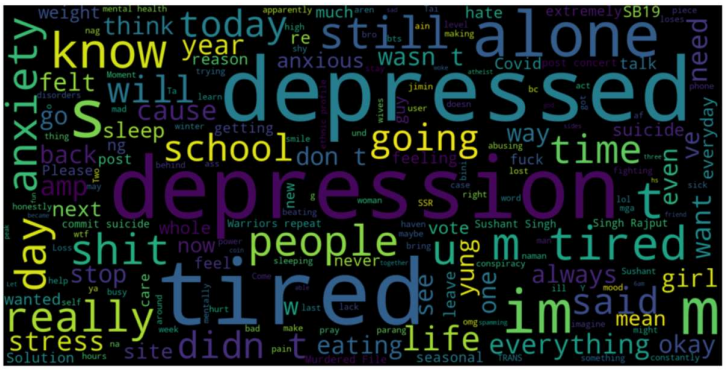



Figure 4.4: Example of Negative Sentiment Wordcloud

xi. Bar plot of word frequency

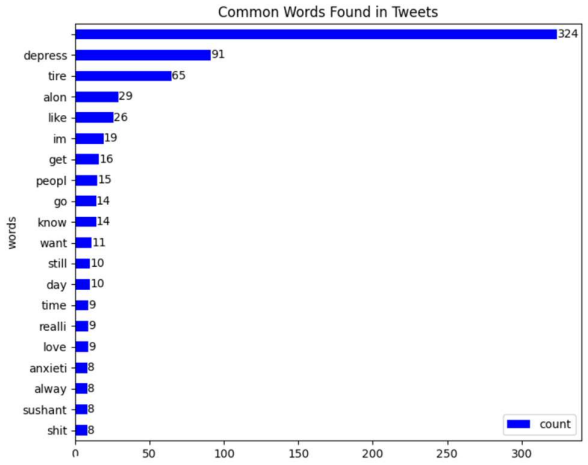


Figure 4.5: Example of Bar plot of word frequency

From those illustration, the plotting and results and posted to a dashboard in the anvil server so it would be illustrate better as well gathered in one board.

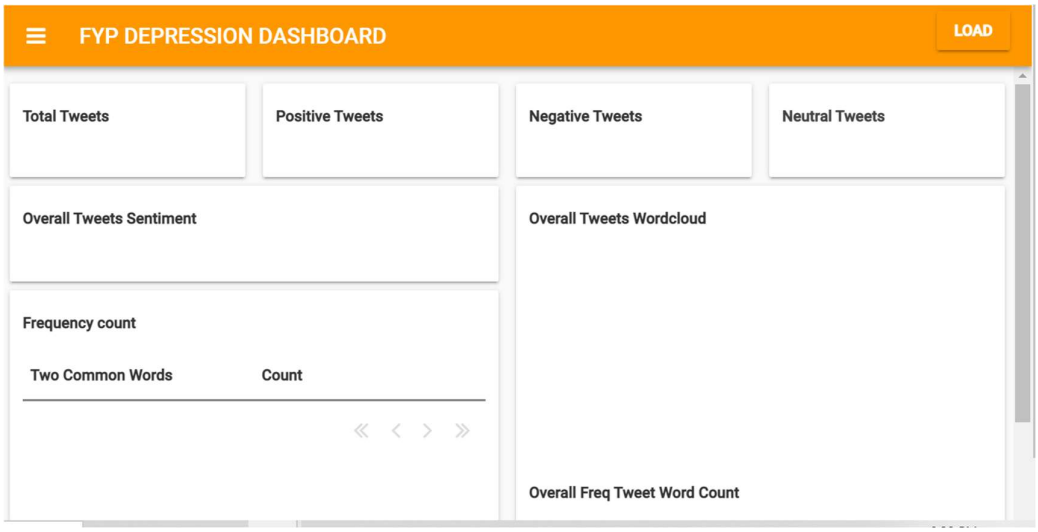



Figure 4.6: Initial Screen of the dashboard

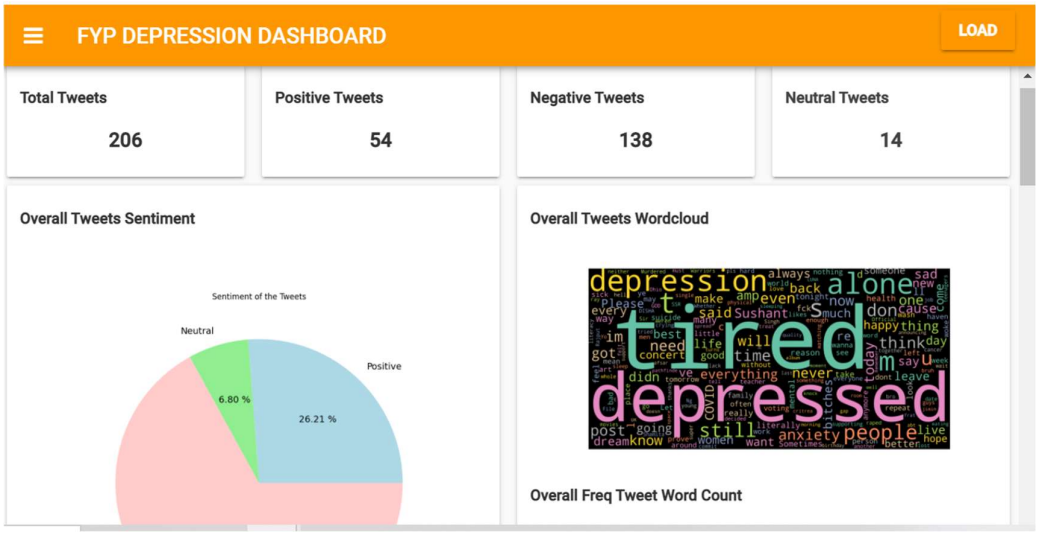



Figure 4.7: Result Screen of the dashboard (1)


Figure 4.8: Result Screen of the dashboard (2)

The plots are now placed orderly and more organize in the dashboard as it can be expressed more easily and understood. All the plots show how sentiment analysis could measure the sentiment from the tweets that was extracted and later on displayed to give a more significant meaning to the study.

## 4.2 DISCUSSION

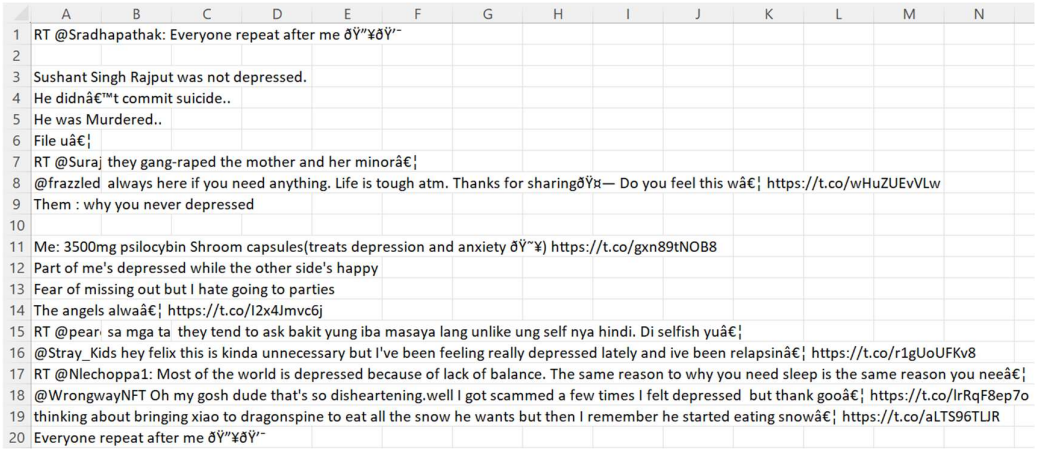
There were several processing methods that was used in the project coding, these methods are all crucial to help to achieve a better end result for this project. Several processing methods that will be discussed here are:

i. Filtration on tweets

For the filtration process it is necessary to dispose of all the unrequired messages that are in the tweets. The tweets are initially subjected to many things that was remove such as url, emoji, retweets, image or videos and many more unrelatable things that aren't necessary for the tweet's sentiment. The code

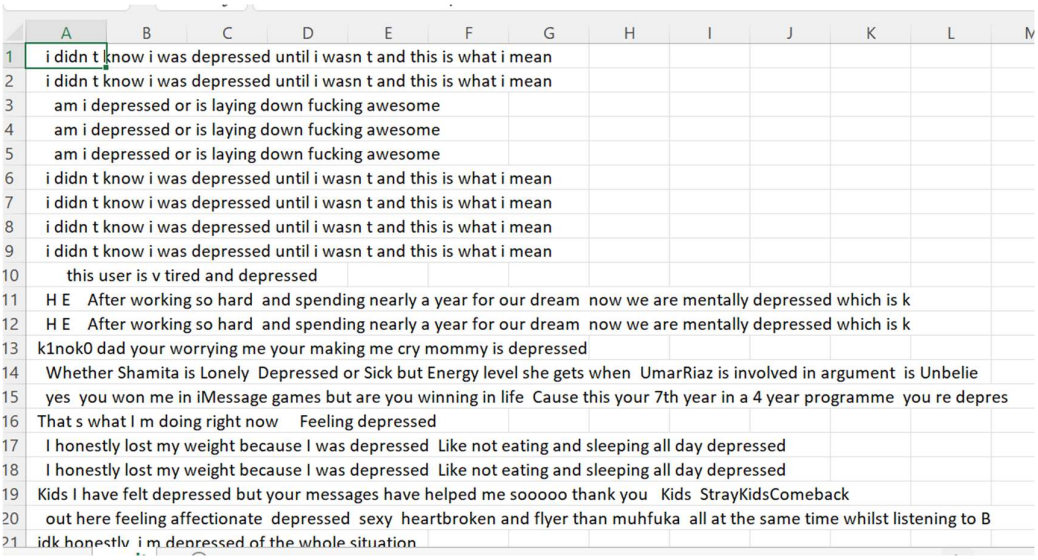
```
line = re.sub("(@[A-Za-z0-9]+)|([^0-9A-Za-z
\t])|(\w+:\/\/\S+)|(#)|(RT[\s]+)|(https?:\/\/\S+)", " ", tweet.text)
```

is used to address this matter and remove all. Below is the example of a before and after filtration on the tweets:



| | A | B | C | D | E | F | G | H | I | J | K | L | M | N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | RT @Sradhapathak: Everyone repeat after me ðŸ™¥ðŸ˙ | | | | | | | | | | | | | |
| 2 | | | | | | | | | | | | | | |
| 3 | Sushant Singh Rajput was not depressed. | | | | | | | | | | | | | |
| 4 | He didnâ€™t commit suicide.. | | | | | | | | | | | | | |
| 5 | He was Murdered.. | | | | | | | | | | | | | |
| 6 | File uâ€¦ | | | | | | | | | | | | | |
| 7 | RT @Suraj they gang-raped the mother and her minorâ€¦ | | | | | | | | | | | | | |
| 8 | @frazzled always here if you need anything. Life is tough atm. Thanks for sharingðŸ¤— Do you feel this wâ€¦ https://t.co/wHuZUEvVLw | | | | | | | | | | | | | |
| 9 | Them : why you never depressed | | | | | | | | | | | | | |
| 10 | | | | | | | | | | | | | | |
| 11 | Me: 3500mg psilocybin Shroom capsules(treats depression and anxiety ðŸ˜¥) https://t.co/gxn89tNOB8 | | | | | | | | | | | | | |
| 12 | Part of me's depressed while the other side's happy | | | | | | | | | | | | | |
| 13 | Fear of missing out but I hate going to parties | | | | | | | | | | | | | |
| 14 | The angels alwaâ€¦ https://t.co/I2x4Jmvc6j | | | | | | | | | | | | | |
| 15 | RT @pear sa mga ta they tend to ask bakit yung iba masaya lang unlike ung self nya hindi. Di selfish yuâ€¦ | | | | | | | | | | | | | |
| 16 | @Stray_Kids hey felix this is kinda unnecessary but I've been feeling really depressed lately and ive been relapsinâ€¦ https://t.co/r1gUoUFKv8 | | | | | | | | | | | | | |
| 17 | RT @Nlechoppa1: Most of the world is depressed because of lack of balance. The same reason to why you need sleep is the same reason you neeâ€¦ | | | | | | | | | | | | | |
| 18 | @WrongwayNFT Oh my gosh dude that's so disheartening.well I got scammed a few times I felt depressed but thank gooâ€¦ https://t.co/lrRqF8ep7o | | | | | | | | | | | | | |
| 19 | thinking about bringing xiao to dragonspine to eat all the snow he wants but then I remember he started eating snowâ€¦ https://t.co/aLTS96TLJR | | | | | | | | | | | | | |
| 20 | Everyone repeat after me ðŸ™¥ðŸ˙ | | | | | | | | | | | | | |

Figure 4.9: Example of before filtration process



| | A | B | C | D | E | F | G | H | I | J | K | L | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | i didn t know i was depressed until i wasn t and this is what i mean | | | | | | | | | | | | |
| 2 | i didn t know i was depressed until i wasn t and this is what i mean | | | | | | | | | | | | |
| 3 | am i depressed or is laying down fucking awesome | | | | | | | | | | | | |
| 4 | am i depressed or is laying down fucking awesome | | | | | | | | | | | | |
| 5 | am i depressed or is laying down fucking awesome | | | | | | | | | | | | |
| 6 | i didn t know i was depressed until i wasn t and this is what i mean | | | | | | | | | | | | |
| 7 | i didn t know i was depressed until i wasn t and this is what i mean | | | | | | | | | | | | |
| 8 | i didn t know i was depressed until i wasn t and this is what i mean | | | | | | | | | | | | |
| 9 | i didn t know i was depressed until i wasn t and this is what i mean | | | | | | | | | | | | |
| 10 | this user is v tired and depressed | | | | | | | | | | | | |
| 11 | H E After working so hard and spending nearly a year for our dream now we are mentally depressed which is k | | | | | | | | | | | | |
| 12 | H E After working so hard and spending nearly a year for our dream now we are mentally depressed which is k | | | | | | | | | | | | |
| 13 | k1nok0 dad your worrying me your making me cry mommy is depressed | | | | | | | | | | | | |
| 14 | Whether Shamita is Lonely Depressed or Sick but Energy level she gets when UmarRiaz is involved in argument is Unbelie | | | | | | | | | | | | |
| 15 | yes you won me in iMessage games but are you winning in life Cause this your 7th year in a 4 year programme you re depres | | | | | | | | | | | | |
| 16 | That s what I m doing right now Feeling depressed | | | | | | | | | | | | |
| 17 | I honestly lost my weight because I was depressed Like not eating and sleeping all day depressed | | | | | | | | | | | | |
| 18 | I honestly lost my weight because I was depressed Like not eating and sleeping all day depressed | | | | | | | | | | | | |
| 19 | Kids I have felt depressed but your messages have helped me sooooo thank you Kids StrayKidsComeback | | | | | | | | | | | | |
| 20 | out here feeling affectionate depressed sexy heartbroken and flyer than muhfuka all at the same time whilst listening to B | | | | | | | | | | | | |
| 21 | idk honestly i m depressed of the whole situation | | | | | | | | | | | | |

Figure 4.10: Example of after filtration process

With the filtration process and the line of code that was mentioned, the data seems more readable and understandable compared to before the filtration process.

ii. Sentiment scoring

The sentiment scoring implements the Sentiment Intensity Analyzer VADER method which have helped me to calculate the scoring of each tweet's sentiment. VADER (Valence Aware Dictionary and Sentiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is tuned in to social media sentiments. VADER utilizes a mix of techniques. A sentiment lexicon is a collection of lexical features (e.g., words) that are classified as positive or negative depending on their semantic orientation.

VADER not only displays the Positivity and Negativity scores, but also the degree to which a sentiment is positive or negative.

```
import nltk
import time
nltk.download('vader_lexicon')
from nltk.sentiment import SentimentIntensityAnalyzer
import operator
sia = SentimentIntensityAnalyzer()
dataset["sentiment_score"] = dataset['Tweets'].apply(lambda x: sia.polarity_scores(x)["compound"])
```

Figure 4.11: VADER lexicon implementation

This method has assisted me in scoring out the sentiments based on its existing algorithm to help measure the sentiment in each of the tweets. The compound score is made up of positive, negative, and neutral scores, which are then adjusted between -1 (very negative) and +1 (very positive).

- There is no need for any training data.
- It can decipher the meaning of a text that includes emoticons, slang, conjunctions, capital letters, punctuation and more.
- It's great for text on social networking.
- VADER has the ability to work with several domains.

Following is the example of the implementation of the sentiment analysis method used to measure sentiment which has been classed to three categories, positive, neutral and negative:

| Index | Tweets | timent_sc | sentiment |
|---|---|---|---|
| 82 | Lol a 2 riddim alone me mek since November enuh | 0.2023 | pos |
| 83 | Kashi Vishwanath corridor is Hindu culture ge… | 0.1027 | pos |
| 84 | sakura was a 48g member iz one member with o… | -0.25 | neg |
| 85 | alone | -0.25 | neg |
| 86 | The Lord puts people who suffer on our path … | -0.1027 | neg |
| 88 | I heard there was 30 increase in deaths at h… | -0.5574 | neg |
| 91 | Sometimes you just need a break In a beauti… | 0.4404 | pos |
| 92 | 11 this is why ur alone in the uk | -0.25 | neg |
| 93 | Waking up home alone hits diff | -0.25 | neg |
| 94 | It was a bad shoot Looked pretty obvious At l… | -0.631 | neg |
| 95 | Never have I felt more isolated and alone while u… | -0.5936 | neg |
| 96 | stuck after 3 years leave me alone mf | -0.4939 | neg |
| 99 | eye NO WAIT COME BACK YOU CAN T LEAVE ME A… | -0.8893 | neg |
| 100 | Let s unite every one let your voice reach o… | 0.212 | pos |
| 101 | i honestly don t mind sitting alone in my noseble… | 0 | neu |
| 102 | When we struggle with our mental wellbeing so… | -0.6705 | neg |
| 103 | Opening different livestreams are like enteri… | -0.4939 | neg |
| 104 | OHIO CANT HANDLE THIS THE BULLY HAS A BROK… | -0.9231 | neg |
| 105 | I was putting lights on the tree alone and … | 0.5106 | pos |

Figure 4.12: Example of Sentiment scoring

iii. Subjectivity and polarity

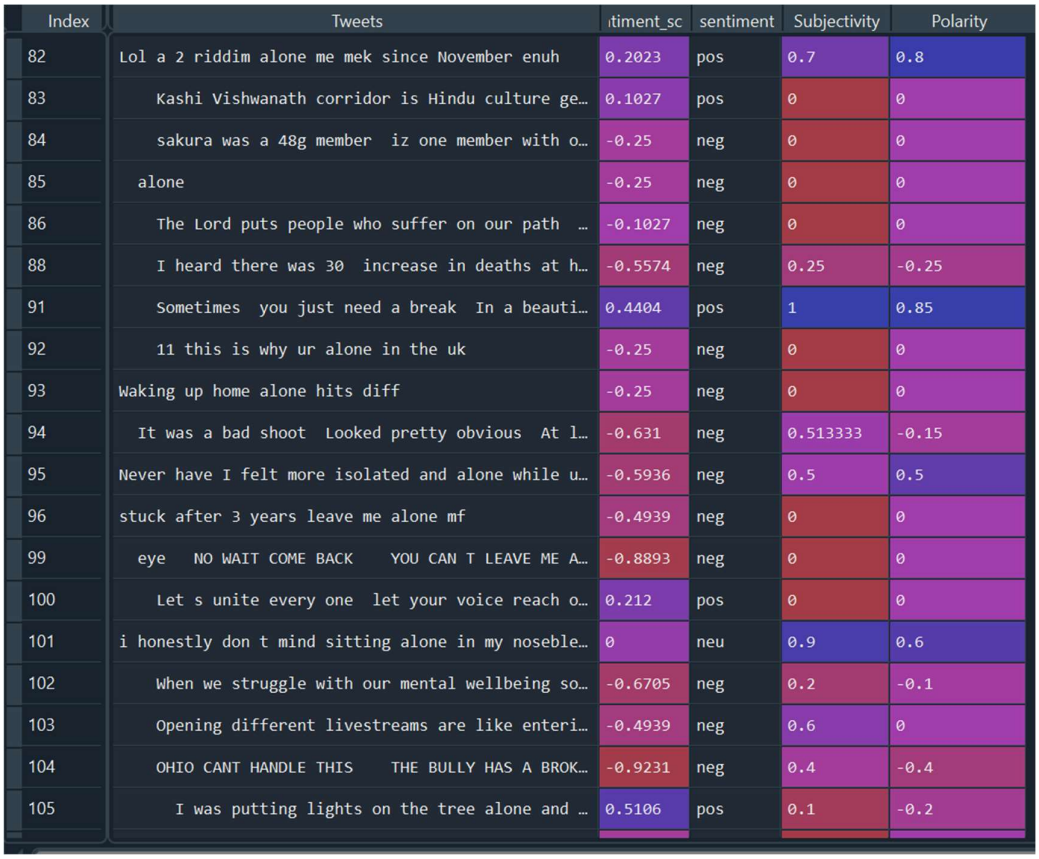
The subjectivity and polarity is also an alternative way to measure the scoring of sentiment which uses Textblob library. TextBlob is a Natural Language Processing (NLP) Python package (NLP). Natural Language ToolKit (NLTK) was used extensively by TextBlob to complete its objectives. NLTK is a library that allows users to work with categorization, classification and a variety of other tasks by providing easy access to a large number of lexical resources. TextBlob is a basic package that allows for extensive textual data analysis and operations.

```
111  from textblob import TextBlob
112
113  def getSubjectivity(text):
114      return TextBlob(text).sentiment.subjectivity
115
116  #Create a function to get the polarity
117  def getPolarity(text):
118      return TextBlob(text).sentiment.polarity
```

Figure 4.13: Textblob implementation

The polarity and subjectivity are the subject returned when using TextBlob. It's range of polarity is [-1,1] where -1 indicate a negative sentiment and 1 indicate a positive sentiment. Negative words can change the polarity of the sentence. TextBlob semantic labels are useful for fine-grained analysis. For example, emojis, exclamation marks, emojis. Between [0,1] is subjectivity. The degree of personal opinion and factual information in a text is measured by subjectivity. Due to the increased subjectivity of the text, it contains more personal opinions than factual information. TextBlob has another setting: Strength. "Intensity" is used by TextBlob to calculate subjectivity. The strength of one word affects whether you change the next word. Adverbs are used as English modifiers ("very good").

Following is the example of the scoring used by textblob which gives out subjectivity and polarity for the tweets on its sentiment:



| Index | Tweets | itiment_sc | sentiment | Subjectivity | Polarity |
|---|---|---|---|---|---|
| 82 | Lol a 2 riddim alone me mek since November enuh | 0.2023 | pos | 0.7 | 0.8 |
| 83 | Kashi Vishwanath corridor is Hindu culture ge… | 0.1027 | pos | 0 | 0 |
| 84 | sakura was a 48g member iz one member with o… | -0.25 | neg | 0 | 0 |
| 85 | alone | -0.25 | neg | 0 | 0 |
| 86 | The Lord puts people who suffer on our path … | -0.1027 | neg | 0 | 0 |
| 88 | I heard there was 30 increase in deaths at h… | -0.5574 | neg | 0.25 | -0.25 |
| 91 | Sometimes you just need a break In a beauti… | 0.4404 | pos | 1 | 0.85 |
| 92 | 11 this is why ur alone in the uk | -0.25 | neg | 0 | 0 |
| 93 | Waking up home alone hits diff | -0.25 | neg | 0 | 0 |
| 94 | It was a bad shoot Looked pretty obvious At l… | -0.631 | neg | 0.513333 | -0.15 |
| 95 | Never have I felt more isolated and alone while u… | -0.5936 | neg | 0.5 | 0.5 |
| 96 | stuck after 3 years leave me alone mf | -0.4939 | neg | 0 | 0 |
| 99 | eye NO WAIT COME BACK YOU CAN T LEAVE ME A… | -0.8893 | neg | 0 | 0 |
| 100 | Let s unite every one let your voice reach o… | 0.212 | pos | 0 | 0 |
| 101 | i honestly don t mind sitting alone in my noseble… | 0 | neu | 0.9 | 0.6 |
| 102 | When we struggle with our mental wellbeing so… | -0.6705 | neg | 0.2 | -0.1 |
| 103 | Opening different livestreams are like enteri… | -0.4939 | neg | 0.6 | 0 |
| 104 | OHIO CANT HANDLE THIS THE BULLY HAS A BROK… | -0.9231 | neg | 0.4 | -0.4 |
| 105 | I was putting lights on the tree alone and … | 0.5106 | pos | 0.1 | -0.2 |

Figure 4.14: Example of Textblob implementation

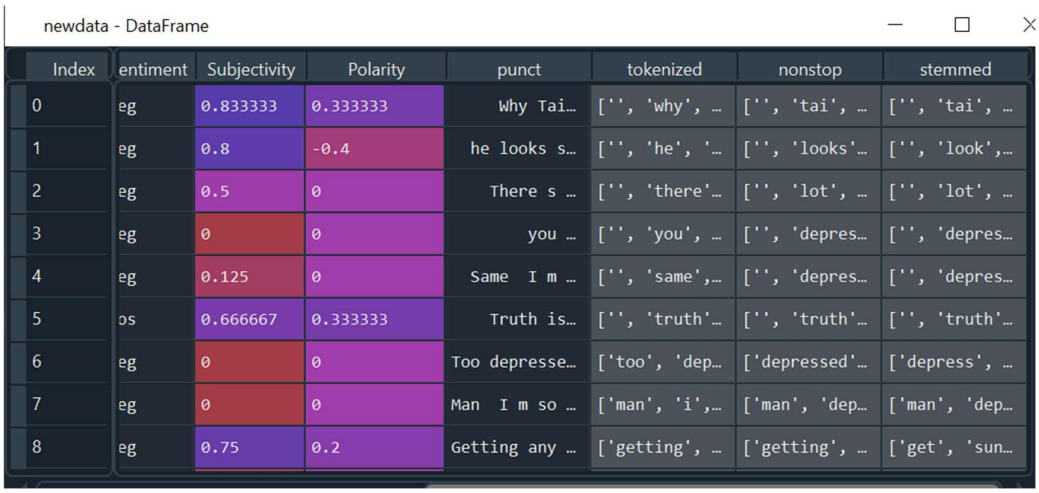
iv. Cleansing process

Cleansing process is as well a necessary process that should be utilized as it helps to eliminate all the irrelevant and unnecessary words from the tweets. This will make it easier to count the frequency of each word from the tweets as common words like “he”, “she”, “them”, “is” (removing stopwords) as well rooting out words such as “depression”, “depressed” to “depress” (applying stemming).

newdata - DataFrame

| Index | Tweets | itiment_sc | sentiment | Subjectivity | Polarity |
|---|---|---|---|---|---|
| 0 | Why Tai wanted to prove SSR is depressed and… | -0.0516 | neg | 0.833333 | 0.333333 |
| 1 | he looks so fucking depressed holy shit | -0.8287 | neg | 0.8 | -0.4 |
| 2 | There s a lot of go for a walk and you don… | -0.5719 | neg | 0.5 | 0 |
| 3 | you depressed mf | -0.5106 | neg | 0 | 0 |
| 4 | Same I m so depressed | -0.5563 | neg | 0.125 | 0 |
| 5 | Truth is I am super depressed It pains me t… | 0.7783 | pos | 0.666667 | 0.333333 |
| 6 | Too depressed to drink I cut myself off | -0.6597 | neg | 0 | 0 |
| 7 | Man I m so depressed The world will probably lo… | -0.5563 | neg | 0 | 0 |
| 8 | Getting any sunlight Your skin absorbs important… | -0.5859 | neg | 0.75 | 0.2 |

Figure 4.15: Example of before cleansing process




| Index | entiment | Subjectivity | Polarity | punct | tokenized | nonstop | stemmed |
|---|---|---|---|---|---|---|---|
| 0 | eg | 0.833333 | 0.333333 | Why Tai… | ['', 'why', … | ['', 'tai', … | ['', 'tai', … |
| 1 | eg | 0.8 | -0.4 | he looks s… | ['', 'he', '… | ['', 'looks'… | ['', 'look',… |
| 2 | eg | 0.5 | 0 | There s … | ['', 'there'… | ['', 'lot', … | ['', 'lot', … |
| 3 | eg | 0 | 0 | you … | ['', 'you', … | ['', 'depres… | ['', 'depres… |
| 4 | eg | 0.125 | 0 | Same I m … | ['', 'same',… | ['', 'depres… | ['', 'depres… |
| 5 | os | 0.666667 | 0.333333 | Truth is… | ['', 'truth'… | ['', 'truth'… | ['', 'truth'… |
| 6 | eg | 0 | 0 | Too depresse… | ['too', 'dep… | ['depressed'… | ['depress', … |
| 7 | eg | 0 | 0 | Man I m so … | ['man', 'i',… | ['man', 'dep… | ['man', 'dep… |
| 8 | eg | 0.75 | 0.2 | Getting any … | ['getting', … | ['getting', … | ['get', 'sun… |

Figure 4.16: Example of after cleansing process

By having all these cleansing process, it would provide a better measure and evaluation on how each word are considered and taken into consideration for the frequency counting.

# CHAPTER 5

# CONCLUSION AND RECOMMENDATION

## 5.1 CONCLUSION

In conclusion, depression is now addressed as a star topic as it is a serious mental illness that can severely affect a person's life in many ways possible. It is crucial to analyse the sentiment expressiveness that is in Twitter. In regard to that the sentiment analysis will help identify the sentiment of depression in tweets. With the assistance of sentiment analysis method where the identification for the depression polarity within the tweets will help determine the rate of intensity of depression where it was mentioned as the second health problem in Malaysia. This method will help me assess live dataset from Twitter and visualize the gathered data to validate the statement that was announced by UM Specialist Centre. This project will help prove that mental illness is actually a serious matter as well as an actual health problem that is suffered by a huge percentage of Malaysian despite their age.

The preliminary studies that had been done by research and literature reviews had provided some insight and deeper understanding of how this technique could help in the studies to get data for depression from social media. It is answered that the technique can be used as a way to measure depression based on a person's emotional integrity, which can be defined, correlated with data and examined (Stephen & Prabu, 2019). It is proven from the research and studies done by other people based on the literature review that machine learning algorithms could be used to diagnose depression among the social media users (Azam et al., 2021). So, with that the studies further on using the sentiment analysis approach to measure and determine the rate of depression within the Malaysian twitter users. The tweets is retrieved based on the allocated queries (location and keywords) and then processed through filtration before measuring the sentiment of each of the tweets.

The clean tweets are then processed through sentiment analysis lexicon-based approach to determine its polarity whether the tweet that was 'expressed' holds positive, neutral or negative sentiments towards emotions and feelings. Based on the

findings it is proven that Twitter had been widely used to expressed oneself and show how they feel about a certain topic. Throughout this project, I have done countless try and error as well deployment to identify working and non-working codes, while at it I have observed that most of the time the percentage of negative sentiment is always higher compared to the percentage of the positive and neutral sentiments. This have proven that numbers of Malaysian are facing depression and openly sharing it to Twitter. This can help conclude that what had been issued by UM Specialist Centre is true, where a lot of Malaysian people are suffering mental health issue.

## 5.2 ACHIEVEMENT OF OBJECTIVES

Based on the objectives that was presented at Chapter 1, it have been fulfil and met as in terms with the aim of this research and project:

- To propose sentiment analysis techniques to analyse sentiment express in tweets with regards to depression

  Sentiment Analysis is used to explore the scope of study that was identified and later on extract the sentiment source known as 'Tweets'. The tweets have been measured and calculated on its polarity which have help determine the expressiveness of the tweet whether it is positive, neutral or negative. Sentiment Analysis have help to analyse the 'emotions' that is portrayed by the sentence written by the users as a whole following the keywords and query condition allocated.

- To validate the ability of the proposed technique in analysing the sentiment leading to depression

  This project have shown the abilities of sentiment analysis where it could measure the sentiments scoring of each tweets that was pulled. The tweets have been specified with query related to depression to help guide it in the term of scoring and understanding the sentiment of the studies. The sentiment of each tweet was then identified and concluded hence concluded that most of the tweets have negative sentiments which is rate as 'depressive emotions'. This have shown that the users are expressing their emotional state to the public in Twitter and that mental illness could be something rather more common than the number of people

that is officially diagnosed. Following the study, there was also classification of words frequency that is commonly used by the individuals that have relations to depression. After numerous executions while developing the project, I could conclude that some of the topic or reasons towards depression are 'life', 'loneliness', 'people' and 'stress'. The tweets studies have shown that the individuals are facing depression based on how their life is, being alone or feeling left behind, the expectation/perspective of people and society towards them and daily dose of pressure or tension events they are facing.

- To develop a dashboard that visualizes the sentiment analysis result into charts and listing

From the backend coding that was made, the coding was the connected to Anvil server where it calls all the functions that was made and connect it to the front end. The visualization that was offered and applicable by Anvil have helped to display the charts and details that we made from the Sentiment Analysis investigation. It has helped to understand the data concluded in an easier and faster way. It also has helped to identify related topics and matters quicker so that it is simpler to assess what related matter that associate to the individuals depressive state.

## 5.3 RECOMMENDATION

From my studies in this project, I have learnt many new things as well approach that could be made to complete and achieve finishing project. The sentiment analysis project that is propose to me by my supervisor, Dr. Ahmad Sobri, had given me the opportunity to explore as well expose myself to further understanding of one of the most important method that is now being used by many industries and organization to help them determine how customers and client sentiment towards something.

Sentiment analysis subject have long existed in the machine learning utilization, yet it is not commonly known or given extra acknowledgment of what it could achieve. The sentiment study could help determine as well make a decision-making process easier as there are many references and subject to be addressed on. This study has shown that sentiment analysis data could be extracted and be utilized into a more meaningful information where it could be used as an illustration. Data are always

something is necessary for any type of development, in addition to that, data visualization helps to make it easier for information to be delivered and received by others.

Recommendation towards this project is that always ensure you understand what you are doing and what will be needed to get the sentiment scoring then dashboarding complete. There are various methods and algorithm that could be used as well as many platforms to utilize for the dashboarding stage, therefore look into which suits you best from your understanding. Having a better understanding of what is needed will help you assess on what's next and necessary. Always to explore and try to identify the best method to be implemented, as some might rule out several lines of codes.

There are many opportunities for improvement and improvise that can be made towards this project such as Test better sentiment calculation methods that could be implemented for this study. As stated in the project activities section, there are many approaches and algorithm that can be used to measure the sentiment of the tweets. Currently the studies is used only to determine the sentiment of the tweets that was pulled, further into the study others can use this to experiment and try out with different algorithm to measure the accuracy, precision and recall of the method used.

Also, future developer and researcher, they could use this study to further construct a better dashboarding interface and visualization for this project. The charts that I have crafted is some of the basics and most common visualization that are normally used for a sentiment analysis studies. Later on, this can be used to find more charts that can be illustrated to help monitor the state of depression among Malaysian Twitter users. The dashboard has help to visualize this project studies and make it simpler and easier for us to determine the proportion of the Malaysian Twitter User on depression, the words that commonly mentioned by each tweets pulled and the relatable words that is in the same tweet.

Other than that, another recommendation that can be given for this project is that others can investigate deeper into the topics of why an individual is facing depression. This can be done when the negative tweets that was classed together extracted and be used to study further on what topic or reason behind their depressive 'confession'. This will help us to determine what actually have made an individual face depression.

**References**


Affairs, D. of V. (1991). Session One - What Causes Mental Illness. *Suppot and Family Education*, 33–36.

American Psychiatric Association. (2020). *What Is Depression?* https://www.psychiatry.org/patients-families/depression/what-is-depression

Azam, F., Agro, M., Sami, M., Abro, M. H., & Dewani, A. (2021). Identifying Depression among Twitter Users using Sentiment Analysis. *2021 International Conference on Artificial Intelligence, ICAI 2021*, *April*, 44–49. https://doi.org/10.1109/ICAI52203.2021.9445271

Higuera, V. (2020). *Everything You Want to Know About Depression*. https://www.healthline.com/health/depression#causes

Hyman, S., Chisholm, D., Kessler, R., Vikram Patel, & Whiteford, H. (2011). Mental disorders. *Voenno-Meditsinskiĭ Zhurnal*, *332*(3), 35–41.

Luo, T., Chen, S., Xu, G., & Zhou, J. (2013). Sentiment Analysis. *Trust-Based Collective View Prediction*, *June 2017*. https://doi.org/10.1007/978-1-4614-7202-5

Mind. (2017). Mental Health Problems (Introduction). *Www.Mind.Org.Uk*, 1–25. https://www.mind.org.uk/information-support/types-of-mental-health-problems/mental-health-problems-introduction/#.XCUs8mT7R1M

Shaikh, R. (2018). *Feature Selection Techniques in Machine Learning with Python*. https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e

Wang, P. S., Aguilar-Gaxiola, S., Alonso, J., Angermeyer, M. C., Borges, G., Bromet, E. J., Bruffaerts, R., de Girolamo, G., de Graaf, R., Gureje, O., Haro, J. M., Karam, E. G., Kessler, R. C., Kovess, V., Lane, M. C., Lee, S., Levinson, D., Ono, Y., Petukhova, M., … Wells, J. E. (2007). Use of mental health services for anxiety, mood, and substance disorders in 17 countries in the WHO world mental health surveys. *Lancet*, *370*(9590), 841–850. https://doi.org/10.1016/S0140-6736(07)61414-7

Yadav, N., Kudale, O., Rao, A., Gupta, S., & Shitole, A. (2021). Twitter Sentiment

Analysis Using Supervised Machine Learning. *Lecture Notes on Data Engineering and Communications Technologies*, *57*(March), 631–642. https://doi.org/10.1007/978-981-15-9509-7_51