UNIVERSITI TEKNOLOGI MARA

A SENTIMENT ANALYSIS APPROACH TO UNIVERSITY STUDENT MENTAL HEALTH

MUHAMMAD AIMAN NAIM BIN MOHD FAIZUL

BACHELOR OF COMPUTER SCIENCE (HONS.)

AUGUST 2023

Universiti Teknologi MARA

A Sentiment Analysis Approach to University Student Mental Health

MUHAMMAD AIMAN NAIM BIN MOHD FAIZUL

Thesis submitted in fulfilment of the requirements for Bachelor of Computer Science (Hons.) College of Computing of Computing, Informatics and Mathematics

August 2023

SUPERVISOR APPROVAL

A SENTIMENT ANALYSIS APPROACH TOWARDS UNIVERSITY STUDENT MENTAL HEALTH

By

MUHAMMAD AIMAN NAIM BIN MOHD FAIZUL
2020490132

This thesis was prepared under the supervision of the project supervisor, Madam Nor Azida Binti Mohamed Noh. It was submitted to the College of Computing, Informatics and Mathematics and was accepted in partial fulfilment of the requirements for the degree of Bachelor of Computer Science (Hons.).

Approved by

…………………………………..

Nor Azida Binti Mohamed Noh
Project Supervisor

JULY 26, 2023

STUDENT DECLARATION

I certify that this thesis and the project to which it refers is the product of my own work and that any idea or quotation from the work of other people, published or otherwise are fully acknowledged in accordance with the standard referring practices of the discipline.

………………………………….
MUHAMMAD AIMAN NAIM
BIN MOHD FAIZUL
2020490132

JULY 26,2023

ACKNOWLEDGEMENT

I would like to express my heartfelt gratitude and appreciation to the following individuals who have played a significant role in the successful completion of this research. First and foremost, I am immensely grateful to Allah for His guidance, blessings, and strength throughout this journey.

I would like to extend my sincere thanks to my supervisor, Madam Nor Azida Binti Mohamed Noh, for her invaluable guidance, support, and expertise. Her insightful feedback, constructive criticism, and encouragement have been instrumental in shaping this research. I am also deeply indebted to my parents for their unwavering love, encouragement, and belief in me. Their constant support and sacrifices have been a constant source of motivation.

Furthermore, I would like to extend my heartfelt appreciation to my dearest friend for their continuous encouragement, understanding, and moral support. Their presence has been a source of inspiration and a reminder that I am not alone on this journey.

I would also like give my special thanks to Nordiana Sahira Binti Norizan for supporting my mental health during this project and also would like to thank all the individuals who have contributed in any way to this research, whether through their participation, advice, or assistance. Your input and collaboration have been invaluable.

Lastly, I would like to acknowledge all the researchers, professionals, and individuals who have contributed to the field of mental health. Your dedication and efforts in understanding and addressing mental health challenges have been a source of inspiration for this research. I am truly grateful to everyone who has contributed to the successful completion of this research. May Allah bless you all abundantly.

ABSTRACT

The prevalence of mental health issues, particularly depression, has increased significantly in recent years. This project aims to address this growing concern by developing a web-based platform that utilizes sentiment analysis techniques to detect signs of depression in user-generated text data. The platform provides a safe and anonymous space for individuals to express their thoughts and feelings while receiving valuable insights and support. The project leverages the power of sentiment analysis algorithms, specifically using Support Vector Machines (SVM) with LinearSVC, to classify user text into depressive or non-depressive sentiments. A comprehensive dataset of randomly collected tweets and expert-approved depression-related texts is used for training and evaluation. The data is pre-processed using techniques such as text cleaning, tokenization, and feature extraction using TF-IDF vectorization. The web-based platform incorporates various key features, including an Anonymous Space where users can freely share their emotions, a Sentiment Analysis page providing real-time insights on the user's mental health based on their input text, and a comprehensive resource section with informative articles on mental health. Rigorous testing has been conducted to validate the system's functionality, usability, and accuracy, yielding impressive results with a 98.97% accuracy rate. Future recommendations include expanding data sources beyond Twitter to enhance representativeness, integrating additional features such as user feedback mechanisms, and collaborating with mental health professionals to provide personalized recommendations and interventions. In summary, this project represents a promising step towards leveraging sentiment analysis and web-based platforms for early detection and support of individuals experiencing depression, with the potential significantly impact on mental health outcomes.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

SVM	Support Vector Machine
SVC	Support Vector Classifier
TF	Term Frequency
IDF	Inverse Document Frequency
NLP	Natural Language Processing
AI	Artificial Intelligence
KNN	K-Nearest neighbour
WHO	World Health Organization
LMIC	Low- and Middle-Income Countries
SA	Sentiment Analysis
RNN	Recurrent Neural Network
HTML	Hyper Text Markup Language
SLCNN	Single Layer Convolutional Neural Network
ML	Machine Learning
ANN	Artificial Neural Network
LIME	Local Interpretable Model-Agnostic Explanations
SHAP	Shapley Additive Explanations
CSS	Cascading Style Sheets
SDLC	Software Development Life Cycle
IDE	Integrated Development Environment
UML	Unified Modelling Language
SUS	System Usability Scale
TN	True Negative
FP	False Positive
FN	False Negative
TP	True Positive

CHAPTER 1

INTRODUCTION

The background and purpose of the study are explained in this chapter. This chapter provides information on the importance of mental health education websites for college students as well as the concerns and issues that motivated this study.

1.1 Background of the study

Today the entire world is witnessing stress and its consequences. Stress becomes an integral part of human life. Anything that creates a challenge or a threat to our comfort is a stress. In academics, stress is unavoidable among students, and it influencing them in all academic activities(Subramani & Kadhiravan, 2017) . During the university years, academic stressors may have shown in any respect of the child environment: home, school, neighbourhood, or friendship. The impact of the academic stress is also outcomes in the areas of exercise, nutrition, and self-care(Subramani, Chellamuthu. Subramanian, 2017).

According to a World Health Organization (WHO) survey, over 30% of college students in eight different countries were estimated to have mental distress. This had a significant negative impact on their ability to study and function in daily life, which resulted in role impairments and subpar academic performance(World Health Organization, 2022).

The issue of mental health among university students is a significant concern. The transition to university can be challenging for many students, and they may experience a range of mental health issues such as anxiety, depression, stress, and other related disorders. This issue can have a significant impact on academic performance, social relationships, and overall well-being(Subramani, Chellamuthu. Subramanian, 2017).

Studies has shown at how the COVID-19 epidemic has affected university students' mental health, as many have had their academic and personal life significantly disrupted by the pandemic. The epidemic had a negative impact on 80% of those surveyed, according to data gathered in April 2020 from more than 2000 college students by Active Minds, a national non-profit organisation devoted to educating and raising awareness about mental health issues for college students (Seidel et al., 2020).

Globally, the COVID-19 epidemic has had a significant influence on people's mental health. Lockdowns, social distancing measures, and the economic downturn have all contributed to a sense of uncertainty and isolation that has had a negative impact on mental health. Students have been particularly affected since they have had to adopt new social and learning strategies while still managing the pressure of meeting academic requirements(Hussna et al., 2021).

Additionally, according to the WHO, college students' mental health includes a stable mood, coordinated interpersonal relationships, objective self-evaluation, and psychological adaptation. When college students move away from home to attend universities, they must overcome additional problems on their own as they make the journey from adolescent to adulthood. Being in a transitional stage between childhood and adulthood also have to confront with varied type of stressors. Therefore, understanding how to properly treat the mental health problems plaguing university students is an important task in universities and the society (McGowan et al., 2020).

Many researchers were attempted to explore the relationship between academic stress, mental health and other correlates among students. Subramani & Kadhiravan (2017) reported that academic stress had significant negative correlation with academic achievement and mental health of the adolescents, also academic achievement had significant positive correlation with mental health. Also, Subramani & Kadhiravan (2017) recognized that academic stress as stressors which occurred to various reasons such as too many assignments, competitions with other students, failures and poor relationship with other students or teachers.

In China, mental health education of university students was initiated in. Practical experience, exploration and development in this field have played an important role during the past 30 years in improving the mental health education of university students, optimising their psychological quality, and promoting their comprehensive development as well as their growth. One in four young people globally suffer from mental illness, which has a negative impact on relationships, scholastic performance, employment prospects, and general well-being and frequently lasts throughout childhood(Sanci et al., 2019).

This study used the sentiment analysis method to analyse the emotional tone and attitudes expressed in social media posts which is Twitter. Sentiment Analysis is the computational study of people’s opinions, attitudes, and emotions. Sentiment analysis is a method for identifying people' feelings in data that is expressed as positive, negative, or neutral text. It also refers to the systematic identification of emotions using Natural Language Processing, Text Analysis, Computational Linguistics, and Biometrics (Seidel et al., 2020). Collecting data from Twitter can provide insight into the mental health issues experienced by university students, as well as the effectiveness of interventions and treatments.

Twitter is one of these platforms where users may communicate their feelings and emotions combined with other information, such as their opinions and support for a cause or their ideas on specific subjects. Millions of people use Twitter to post updates about their lives. Also, these tweets can be utilised to assess a person's intuitive mental health(Sadasivuni & Zhang, 2019).

People have been open about their behavioural changes and their thoughts on social networking sites like Twitter. These people revealed the unexpected change in their conduct along with a fast mood change, feeling extremely happy or sad, and parts of their mental health. Due to their busy schedules and limited social interaction, the majority of people experience mental health issues like stress and anxiety on a daily basis(Gupta et al., 2021).

1.2 Problem Statement

The prevalence of mental health issues among university students has been a growing concern in recent years. The internet is a common source of information for students seeking help, but misinformation about mental health on the internet poses a significant challenge for individuals seeking accurate and reliable information(Reavley et al., 2021). To address the growing concern of mental health issues among university students and combat the prevalence of misinformation on the internet, this project aims to provide accurate and reliable mental health resources. One of the key solutions that this project will implement is a visually appealing section on the mental health website that showcases a selection of featured articles. This section will highlight the wealth of mental health information available on the platform and encourages users to explore further. Each article covers various aspects of mental health, such as recognizing symptoms, coping strategies, stress management techniques, and personal stories of individuals who have overcome mental health and depression challenges.

Another problem statement related to this project is the stigma surrounding mental health issues. Stigma can prevent individuals from seeking help and can lead to negative attitudes and discrimination towards those with mental health issue(Stuart, 2016). Those who suffer from mental illnesses require a safe place to vent their feelings. They will search for locations where they can express their emotions without fear of social stigma. People can choose from a variety of mediums to express themselves online in current digital age, especially through social media, due to a number of factors(Herdiansyah et al., 2023).

This project aims to reduce the stigma surrounding mental health by providing a safe and anonymous space for students to express their thoughts and feelings about mental health. By empowering students to speak openly about their mental health problems through student-run mental health awareness, this project aims to reduce the stigma surrounding mental health issues and encourage individuals to seek help when needed(Stuart, 2016).

One more problem statement related to the creation of this project is the lack of access to mental health resources for university students. Many universities have limited mental health resources, which can make it difficult for students to access the help they need. This problem can be exacerbated by the high cost of mental health services and the stigma surrounding mental health issues(Sutherland, 2018). According to the Sutherland (2018), many universities have limited mental health resources, which can make it difficult for students to access the help they need. The lack of access to mental health resources can also have long-term consequences for students, affecting their future employment, earning potential, and overall health.

1.3 Project Objective

The objectives of this project are as below:

1. To design a mental health website that are provides a safe and anonymous space for users to express their thoughts and feelings.
2. To develop a sentiment-analysis algorithm using Support Vector Machine (SVM) to determine whether a user is exhibiting symptoms of a depression based on their input text.
3. To test the functionality, usability, and the accuracy of the website system.

1.4 Scope

In this project, the crucial challenge of mental health and its significant effects on people, communities, and society as a whole is taken. The scope of this report revolves around three essential aspects: understanding the target audience, exploring the scope of mental health, and highlighting key features of the proposed system.

1.4.1 Target Audience

This project will be focusing on the university student which is the young people at the age between 19 and 24 years, living in Malaysia, and had sufficient English. Focusing on university students for this project is supported by the fact that student mental health is in crisis, and by nearly every metric, student mental health is worsening (Abrams, 2022).

Additionally, mental health problems can affect many areas of students' lives, reducing their quality of life, academic achievement, physical health, and satisfaction with the college experience, and negatively impacting relationships with friends and family members (Abrams, 2022).

By focusing on university students, this project aims to address the unique mental health challenges that this population faces and provide a targeted solution to help them overcome the barriers to accessing mental health resources.

1.4.2 Mental Health Scope

After doing some research, this project will only be focusing on detecting depression from the user text. Several studies have focused on using natural language processing (NLP) techniques to detect depression from text, including social media posts, clinical interviews, and other sources(Chiong et al., 2021). These studies have shown promising results, with some achieving high accuracy rates in detecting depression from text(Trotzek et al., 2020).

Focusing on detecting depression from user text may also be more feasible and less invasive than other methods, such as clinical interviews or physiological measures(Chiong et al., 2021). Therefore, it seems reasonable to focus on this aspect of mental health detection, given the existing research and the potential benefits of this approach.

1.4.3 System Features

This project consists of several key features that collectively aim to address mental health concerns among university students. The first feature is the Sentiment Analyzer, which allows users to input their text and receive sentiment analysis results.

This feature enables users to input their text, such as social media posts or personal messages, and receive results indicating the presence or absence of depression. By utilizing natural language processing techniques and sentiment analysis algorithms, the system can identify patterns, keywords, and linguistic cues associated with depressive sentiments.

The second feature is the Mental Health Resources section, which serves as a comprehensive information hub. It offers tips, strategies, and coping mechanisms specifically tailored to the challenges faced by university students.

This section covers various mental health problems, such as depression, anxiety, and stress management, empowering users with knowledge and resources to navigate these issues. Additionally, the system provides symptom descriptions and curated videos that educate users about mental health, fostering awareness and understanding.

The third feature is the Dashboard, which presents statistical insights derived from a dataset comprising user text from Twitter and a trusted source-verified dataset of confirmed depression problems. The Dashboard provides visual representations of trends, patterns, and prevalence rates of mental health issues among university students.

The data collection has been collected from the Twitter, comprising user text, as well as a dataset from trusted sources that have confirmed the presence of depression problems in the text. These data-driven insights enable researchers, administrators, and other stakeholders to make informed decisions and implement targeted interventions to improve mental health support systems in universities.

By integrating these features, this project aims to leverage sentiment analysis to detect depression among university students. Through the Sentiment Analyzer, users can gain an understanding of their mental well-being, while the Mental Health Resources section provides guidance and support.

The Dashboard feature showcases the statistical analysis of the dataset, enabling a broader perspective on mental health concerns in the university community. By combining these features, our system aims to contribute to the overall mental well-being of university students and facilitate early intervention for those experiencing depression.

1.5 Project Significance

Mental health is a significant issue that affects many individuals, including university students. Research has shown that mental health problems can impact many areas of students' lives, reducing their quality of life, academic achievement, physical health, and satisfaction with the college experience, and negatively impacting relationships with friends and family members(Gibbon et al., 2020).

Therefore, it is essential to address mental health challenges among university students to improve their overall well-being and academic success(Gibbon et al., 2020). Additionally, providing support for child and student social, emotional, behavioural, and mental health needs can help students overcome the barriers to accessing mental health resources and improve their academic outcomes(Cardona, 2021).

By addressing mental health challenges among university students, this project has the potential to improve the overall well-being and academic success of this population, contributing to a healthier and more productive society.

Reducing the stigma surrounding mental health issues is another significant aspect of this project. Stigma can prevent individuals from seeking help and can lead to negative attitudes and discrimination towards those with mental health issues(Stuart, 2016).

By providing a safe and anonymous space for students to express their thoughts and feelings about mental health, this project aims to reduce the stigma surrounding mental health issues and empower students to seek help when needed.

Additionally, the project aims to provide accurate and reliable information about mental health to university students, helping them make informed decisions about their mental health and reducing the risk of misinformation. By reducing the stigma surrounding mental health issues and providing accessible mental health resources, this project has the potential to improve the overall well-being and academic success of university students, contributing to a healthier and more productive society(Stuart, 2016).

CHAPTER 2

LITERATURE REVIEW

This chapter explore the literature review regarding the study of “A Sentiment Analysis Approach to University Student Mental Health”. In order to fully complete this study, the ideas and methods used in papers, journals, and articles have been examined and analysed. To validate this research, a number of topics that contribute to the project's supporting evidence were also emphasized. The chapter's outline for the literature review of " A Sentiment Analysis Approach to University Student Mental Health " is shown graphically in Figure 2.1.

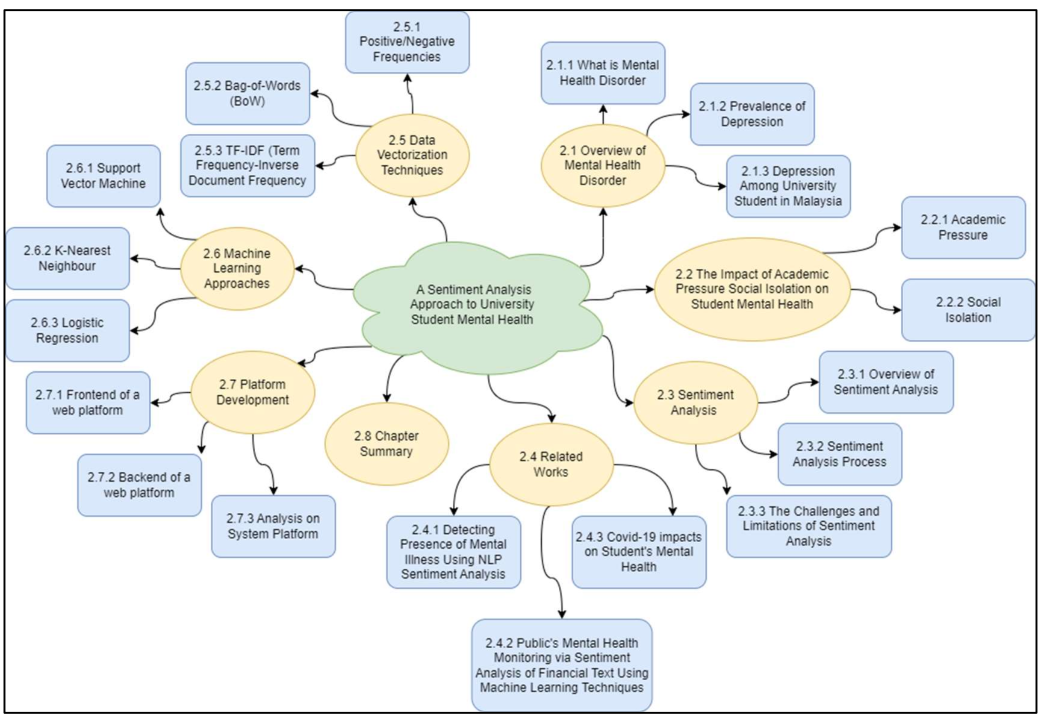


Figure 2. 1 LR Diagram

2.1 Overview of Mental Health Disorder

This section provides an in-depth exploration of mental health disorders, covering three key aspects: understanding what constitutes a mental health disorder, examining the prevalence of depression globally, and focusing on depression's impact specifically among university students in Malaysia. By gaining insights into these crucial aspects, this project aims to recognize the widespread impact of mental health disorders and the need for effective support and interventions to promote mental well-being in individuals and communities.

2.1.1 What is Mental Health Disorder

Mental disorders are characterized by clinically significant disturbances in an individual's cognition, emotional regulation, or behaviour. These disorders are usually associated with distress or impairment in important areas of functioning, such as work, school, or relationships. Mental disorders can take many different forms, including depression, anxiety, bipolar disorder, schizophrenia, and personality disorders(World Health Organization, 2022).

Mental disorders may also be referred to as mental health conditions, which is a broader term covering mental disorders, psychosocial disabilities, and other mental states associated with significant distress or impairment. Mental disorders can be caused by a variety of factors, including genetics, environment, and lifestyle. Exposure to unfavourable social, economic, geopolitical, and environmental circumstances, including poverty, violence, inequality, and environmental deprivation, can also increase people's risk of experiencing mental health conditions(World Health Organization, 2022).

Understanding the scope and impact of mental disorders is crucial to address the mental health needs of individuals effectively. Mental disorders can have significant negative outcomes both academically and psychologically, including difficulty sleeping, changes in weight, and an increased likelihood of suicidal thoughts, in addition to problems with academic learning(Sutherland, 2018).

Mental health systems have not yet adequately responded to the needs of people with mental disorders and are significantly under-resourced. The gap between the need for treatment and its provision is wide all over the world and is often poor in quality when delivered(World Health Organization, 2022).

2.1.2 Prevalence of Depression

The term "depression" refers to a mental health disorder. It might be characterised as sadness, loss, or anger that interferes with daily tasks. Also, it's very common. According to data from the Centres for Disease Control and Prevention Trusted Source, 18.5% of American adults experienced depressive symptoms at some point during a two-week period in 2019 (Villarroel & Terlizzi, 2020).

The prevalence of depression among university students is a significant concern, and research has shown that it is a widespread problem. According to a systematic review and meta-analysis study published in 2020, it found that the overall prevalence of depression among Chinese university students was 28.4%(Gao et al., 2020). The overall prevalence of depressive symptoms among university students in low and middle-income countries (LMICs) was 24.4%, according to a systematic review and meta-analysis study published in 2020(Akhtar et al., 2020).

Depression can have significant negative outcomes both academically and psychologically, including difficulty sleeping, changes in weight, and an increased likelihood of suicidal thoughts, in addition to problems with academic learning(Cassady et al., 2019).

Additionally, mental health problems can affect many areas of students' lives, reducing their quality of life, academic achievement, physical health, and satisfaction with the college experience, and negatively impacting relationships with friends and family members(Croock et al., 2023). These issues can also have long-term consequences for students, affecting their future employment, earning potential, and overall health(Croock et al., 2023).

2.1.3 Depression among University Student in Malaysia

Depression is a significant mental health issue among university students in Malaysia, with research conducted between 2017 and 2023 indicating a high and concerning prevalence. Studies conducted during this period reported rates ranging from 27.5% to 29.3% in a pilot study, 41.6% among health science students, and 20% to 27.5% with moderate depressive symptoms in a repeated cross-sectional study(Yap et al., 2021).

Several key factors contribute to depression in this population. Financial problems, including the high cost of education and living expenses, have been identified as significant stressors leading to depression. Academic pressure, such as the demand for high performance and meeting deadlines, also contributes to increased stress and anxiety(Ashraful Islam et al., 2018). Social isolation, characterized by a lack of social support and difficulty in forming connections, is another factor contributing to depression among university students(Ashraful Islam et al., 2018).

Moreover, the COVID-19 pandemic has exacerbated these issues, causing disruptions to students' lives, social isolation, and financial difficulties. Lastly, the lack of accessible mental health resources further compounds the challenges faced by university students in Malaysia. Addressing these factors and improving mental health support is crucial to effectively tackle the prevalence of depression in this population(Ashraful Islam et al., 2018).

2.2 The Impact of Academic Pressure and Social Isolation on Student Mental Health

This section explores a vital component of mental health: the significant effects of academic pressure and social isolation on students. For many young individuals, the academic journey is marked by intense challenges and growing responsibilities, which can significantly affect their mental well-being. Additionally, the experience of social isolation can further exacerbate these pressures, leading to detrimental effects on their overall mental health.

2.2.1 Academic Pressure

The COVID-19 pandemic has resulted in reduced academic performance among university students due to various factors, including insufficient technological knowledge, support, and equipment for online classes, difficulty concentrating on lectures and understanding teaching materials, and the negative impact of pandemic austerity measures on their lifestyles and psychological well-being. The pandemic also disrupted students' regular routines, which may have negatively affected their psychological well-being and academic performance(Kokkinos et al., 2022).

Financial stress and instability were also identified as significant factors affecting students' life satisfaction and quality of life. The pandemic caused many families to lose jobs or experience reduced salaries, which made it difficult for students to support themselves and their education. This financial stress was found to be the second most influential factor affecting students' overall contentment with their life circumstances. (Kokkinos et al., 2022).

The reduction in life satisfaction among students is not surprising, given that they are considered more vulnerable to external conditions that can impact their subjective well-being, particularly in the cognitive dimension of life satisfaction. The academic pressure experienced by university students is further exacerbated by the challenges brought about by the pandemic, which negatively affect their academic performance and psychological well-being(Kokkinos et al., 2022).

2.2.2 Social Isolation

Prommas et al., (2023) stated that social support is crucial for maintaining good mental health. A lack of social support can lead to feelings of loneliness, isolation, and depression. When people have access to social support, they are more likely to feel a sense of belonging and connection, which can promote better mental health outcomes.

In addition, Prommas et al., (2023) also have found that having strong social support networks can help individuals cope with stress more effectively. When someone is going through a difficult time, having supportive friends or family members to talk to and lean on can make all the difference. This can help to reduce the negative impact of stress on mental health and promote resilience.

Social isolation can have a significant impact on the mental health of university students. Recent studies conducted between 2019 and 2023 have shown some of the consequences. One of the main effects is the higher likelihood of experiencing anxiety and depression among socially isolated students. They tend to show more symptoms of these mental health issues compared to those who have more social connections(Keshavarzi et al., 2021).

Being socially isolated can lead to lower self-esteem and feelings of worthlessness among university students. Social interaction plays a crucial role in shaping our sense of identity and belonging. When students are isolated, they may start to feel like they are not good enough or that they don't matter, which can negatively impact their self-perception and confidence(Keshavarzi et al., 2021).
Another significant impact is the increased risk of turning to substance abuse as a way to cope with the negative emotions that come with isolation. Students who feel socially isolated are more likely to use alcohol and drugs as a means of dealing with their feelings. Research published in 2020 has shown this connection between social isolation and substance use among university students(Hamza et al., 2021).

Tragically, social isolation also raises the risk of suicidal thoughts. The feelings of hopelessness and despair that often accompany isolation can lead to contemplating suicide. A study conducted in 2021 found that socially isolated students were more likely to report having considered taking their own lives(Leal Filho et al., 2021).

To sum up, academic isolation has various consequences for the mental health of university students. These include an increased risk of anxiety, depression, and suicidal thoughts, as well as lower self-esteem and a higher likelihood of turning to substance abuse. It is important for universities and society as a whole to recognize these detrimental effects and develop strategies and support systems to help students combat the negative impacts of social isolation.

2.3 Sentiment Analysis

This section explores into the fascinating field of sentiment analysis, a powerful tool in understanding the emotions and opinions expressed in textual data. Sentiment analysis plays a pivotal role in uncovering valuable insights from vast amounts of information, enabling us to gauge people's sentiments, attitudes, and reactions towards various subjects. By providing an overview of sentiment analysis, exploring its underlying process, and discussing the challenges and limitations it entails, this project aims to shed light on this essential technique's significance and its application in understanding human sentiment within textual content.

2.3.1 Overview of Sentiment Analysis

Sentiment analysis (SA), also known as opinion mining, is a field of study that tries to examine how people feel or perceive many types of entities, including ideas, people, events, problems, services, goods, organisations, and their characteristics. Due to the widespread use of social media and the ease with which messages may be posted, sentiments or opinions from social media offer the most current and comprehensive information.(Yue et al., 2019).

Digital text is subjected to sentiment analysis in an effort to extract human emotions. According to Yadav (2022), processes including text feature processing, user opinion extraction, and sentiment opinion analysis are typically included in sentiment analysis.

The difficult part of sentiment analysis is training and testing machine learning algorithms to analyse the many grammatical phrases, cultural variations, slang, and misspellings that can be heard in word pronunciation. Systems for sentiment analysis might be automatic, rule-based, or hybrid. The method used for the rule-based system is Natural Language Processing. (NLP). As a result, traditional NLP techniques like tokenization, stemming, lemming, and part-of-speech analysis are often used(Rahman et al., 2022).

2.3.2 Sentiment Analysis Process

Sentiment analysis techniques play a crucial role in determining the sentiment expressed in textual data. The sentiment analysis process typically includes the following steps:

1. Data collection: The first step is to gather the text data that wants to be analyze. This could be customer reviews, social media posts, surveys, or any other form of text that contains opinions or sentiments.
2. Text preprocessing: Once the data is collected, it needs to be pre-processed to remove noise and prepare it for analysis. This step involves tasks such as removing special characters and punctuation, converting text to lowercase, and handling contractions or abbreviations.
3. Tokenization: Tokenization involves breaking down the text into individual words or tokens. This step is important for further analysis as it allows the sentiment analysis model to work with discrete units of text.
4. Stop word removal: Stop words are common words that do not carry significant meaning in sentiment analysis, such as "the," "is," or "and." Removing stop words helps reduce noise and improve the efficiency of the analysis.
5. Feature extraction: In this step, relevant features are extracted from the text data. These features could be words, phrases, or even grammatical patterns that are indicative of a particular sentiment. For example,

words like "happy" or "disappointed" might be features that suggest positive or negative sentiment, respectively.

6. Sentiment classification: Once the features are extracted, a machine learning model is trained to classify the sentiment of the text. Various algorithms can be used for sentiment classification, including Naive Bayes, Support Vector Machines (SVM), or Recurrent Neural Networks (RNNs). The model is trained using labelled data where each text sample is associated with a sentiment label.
7. Model evaluation: After training the sentiment analysis model, it needs to be evaluated to assess its performance and accuracy. This is typically done by using a separate test dataset that the model has not seen during training. Evaluation metrics such as accuracy, precision, recall, and F1 score are commonly used to measure the model's effectiveness.
8. Sentiment analysis application: Once the model is trained and evaluated, it can be applied to new, unseen text data to predict the sentiment. The model assigns a sentiment label (positive, negative, or neutral) to each text sample, enabling sentiment analysis on a larger scale.

2.3.3 The Challenges and Limitations of Sentiment Analysis

The exponential growth of social media data has provided new opportunities to extract insights for individuals, business, and governments. However, the quality of the data is crucial for effective sentiment analysis, and challenges arise in both data collection and pre-processing. Data volume poses a challenge during data collection, while data imbalance and varying comment lengths are often encountered in working with primary data(Xu et al., 2022).

Pre-processing the data can be challenging as well, as data from social networking platforms is unstructured, contains noise, inefficiencies, and irrelevant information, and lacks textual emotional content. To extract useful information from such vast amounts of data, a range of data pre-processing techniques is required to turn the text into a predictable and analysable form. Common pre-processing steps include converting uppercase to lowercase, removing unnecessary and multiple pronunciations, removing URL’s, correcting text, and removing stop words. Additionally, some studies exclude nonalphabetic terms during pre-processing(Xu et al., 2022).

The difficult part of sentiment analysis is training and testing machine learning algorithms to analyse the many grammatical phrases, cultural variations, slang, and misspellings that can be heard in word pronunciation. Systems for sentiment analysis might be automatic, rule-based, or hybrid. The method used for the rule-based system is Natural Language Processing. (NLP). As a result, traditional NLP techniques like tokenization, stemming, lemming, and part-of-speech analysis are often used(Rahman et al., 2022).

The challenges and limitations of sentiment analysis outlined in here are important to take as they emphasise the difficulties that can arise when analysing social media data like from Twitter for sentiment. These challenges, such as data volume and pre-processing, can impact the accuracy of the sentiment analysis results.

As such, it is very important to carefully to consider the data collection and pre-processing steps to ensure that data is of high quality and relevant to the sentiment analysis task. Additionally, understanding the range of pre-processing techniques available can help to extract useful information from the social media data and turn it into a predictable and analysable form. This can ultimately be helpful to achieve accurate and meaningful results for the sentiment analysis project.

2.4 Related Work

This section talks about the work that is almost related to sentiment analysis and their ways to implement them.

2.4.1 Detecting Presence of Mental Illness Using NLP Sentiment Analysis - (Yadav, 2022)

The first paper that is very much related to this upcoming project is titled Detecting Presence of Mental Illness Using NLP Sentiment Analysis. This paper focus on techniques that are pertinent to mental health, which is defined as a condition of well-being for the mind, body, and society as a whole rather than just the absence of disease. Their purpose of the study is to identify depression in social media users(Yadav, 2022).

This study argues that mental health is essential to one's overall physical, psychological, and social welfare. This research proposes a system for early detection and intervention in assessing a person's mental condition. It explores how to forecast a user's mental condition while still protecting their privacy by using data mining from numerous sources, including social media and mobile devices(Yadav, 2022).

In order to forecast user information like location, mood, activity, mental condition, depression, anxiety, stress, and other things, the article suggests utilising sentiment analysis and natural language processing on the collected data. Analysis of a person's tweets can reveal information about their mental health. This information can be used by the doctors to successfully treat patients. The forensic professionals can utilise this information to ascertain if a suicide was committed or whether a person is suicidal(Yadav, 2022).

The purpose of using natural language processing as well as sentiment analysis machine learning algorithms also known as opinion mining is because it can identify the emotional tone behind online conversations(Yadav, 2022).

The first thing they do is gathering data from the data sources that is from the social media. A person's username, user ID, age, the text that is uploaded, photographs, time, and other information are collected from social media sites in order to analyse the type of post that the user is sharing. HTML or text files are used as the input text format. After the data is extracted, identification and classification are completed, which assist in the analysis of the positive, negative, and neutral turn off the post(Yadav, 2022).

The data are then pre-processed and saved in two datasets with rows and columns for each field of the user data. Once you're done with data collection, then move on to the next part that is medium of analysis or opinion mining. Once the data has been retrieved, it is analysed, and statistics based on the amount of positive, negative, and neutral ratings of the text are produced. If the overall text contains a positive outcome, the user is seen to be in a healthy mental state, and if the text post outcome is negative, the user is thought to have a mental disorder(Yadav, 2022).

After the mining part of the opinion has been completed, there is another last part that has not been done which is analysis by health care professionals. In order to aid in the early identification of mental illness in a person, a medical expert may consider the user for consultation if the text under analysis contains the greatest number of negative tones. Medical experts can assist a patient who is or will be experiencing mental discomfort or depression with the use of this analysis (Yadav, 2022).

A broad concept of how the system would be used is shown in the Figure 2.2 below.

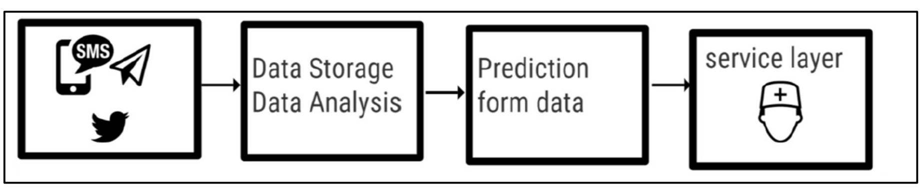


Figure 2. 2 General Flow of the System
(*Source*: Yadav, 2022)

2.4.2 Public's Mental Health Monitoring via Sentimental Analysis of Financial Text Using Machine Learning Techniques - (Alanazi et al., 2022)

The second paper to be discussed is titled Public's Mental Health Monitoring via Sentimental Analysis of Financial Text Using Machine Learning

Techniques. The sentimental analysis of financial text news, which is mostly published on digital media, is one of the main goals of this study. It aims to analyse the public's mental health and determine the effects of national or worldwide financial policies. The dataset was gathered using the Guardian application programming interface, and it was processed using the support vector machine, AdaBoost, and single layer convolutional neural network(Alanazi et al., 2022).

Machine learning approaches can be broadly divided into two categories: supervised learning, which involves the user displays and supplies the learning data, and unsupervised learning, where the learning data is learned using a clustering algorithm while taking the size of the dataset into account. This study's primary challenge is to categorise public attitudes from the enormous textual dataset that is now accessible in order to determine people's general perspectives on financial issues that eventually affect the public's mental health(Alanazi et al., 2022).

The initial goal of this project is to use public datasets linked to financial news items that were acquired using The Guardian API. The dataset based on financial news material is pre-processed for accurate and efficient segmentation into four emotional attributes: neutral, happy, depressed, and irritated, as shown in Figure 2.3.(Alanazi et al., 2022).

Figure 2. 3 Different Emotional States from Circumplex Model
(*Source*: Alanazi et al., 2022)

Their objective is to develop a machine learning-based model for evaluating the intended content of financial news using pre-processed datasets from digital platforms. They will look into how financial news in digital media affects public perception and mental health. Sentimental signals have been used commonly in linguistic representations of target information content concealed in financial literature. With the help of frequency-based descriptors derived from emotive phrases, they develop a baseline to describe financial news content(Alanazi et al., 2022).

The ultimate objective of this project is to develop an accurate machine learning based tool that will support public financial and non-financial organisations. Because then they can handle a high dimensional volumetric dataset, the single layer convolutional neural network (SLCNN), support vector machine (SVM), and AdaBoost are used to categorise gathered datasets related to finance(Alanazi et al., 2022).

In order to rapidly classify the provided text-based dataset into one of four chosen individual attitudes, they compared three machine learning (ML) based techniques, support vector machine (SVM), AdaBoost, and single layer convolutional neural network (SLCNN).

Numerous researchers frequently use these strategies to categorise textual data because to the daily development and expansion of machine learning (ML) techniques. Single layer convolutional neural network (SLCNN) has an accuracy rate of 83.4%, whereas support vector machine (SVM) and AdaBoost have accuracy rates of 57.2% and 66.4%, respectively. Consequently, it is utilised whenever a categorical dependent variable or dependent aim is present.(Alanazi et al., 2022).

2.4.3 Covid-19 impacts on students' Mental Health: Explainable AI and Classifiers - (Hussna et al., 2021)

Now move on to the third paper titled Covid-19 impacts on students' Mental Health: Explainable AI and Classifiers. The article presents a methodology for using machine learning algorithms to predict the impact of COVID-19 on students' mental health. The authors used a dataset that was collected through a web-based survey conducted on the Google Forms platform. The dataset includes responses from 1182 students from various colleges and universities who answered questions related to the pandemic's impact on their social life, mental health, and education(Hussna et al., 2021).

The authors first pre-processed the data, removing missing values, encoding categorical data, labelling the data, and applying standard scaling to the independent variables. They also converted "yes" and "no" responses to numeric values. They then applied several supervised learning algorithms, including Logistic Regression, Random Forest, XGBoost, and AdaBoost, to predict the model's performance. They also trained an Artificial Neural Network (ANN) model and evaluated all the models' accuracy(Hussna et al., 2021).
To make the models more interpretable, the authors used two explainable AI techniques called Local Interpretable Model-Agnostic Explanations (LIME) and Shapley Additive Explanations (SHAP). LIME generates a set of perturbed instances around a given data point and computes the corresponding predictions to explain the original prediction(Hussna et al., 2021).

It then generates a comprehensible model based on the generated data that approximates and explains the original model. SHAP calculates the Shapley values for each specific feature, with each Shapley value representing the impact of the feature in the prediction(Hussna et al., 2021).

Overall, the article provides a useful methodology for predicting the impact of COVID-19 on students' mental health using machine learning algorithms and XAI techniques. The authors' use of multiple algorithms and XAI methods enhances the models' interpretability and trustworthiness, which is crucial for addressing such an important issue.

Table 2.1 below shows the comparison of the previous related works.

Table 2. 1 Comparison of the previous related works

Year	2022	2020	2022
System/app	Detecting Presence of Mental Illness Using NLP Sentiment Analysis	Covid-19 impacts on students' Mental Health: Explainable AI and Classifiers	Public's Mental Health Monitoring via Sentimental Analysis of Financial Text Using Machine Learning Techniques
Objective	To develop a sentiment tool for early detection and prevention of mental health issues in social media users.	To predict the impacts of students' mental health conditions during the COVID-19 pandemic.	Develop a machine learning-based model for evaluating financial news.
Version /Platform	Web-based application	Not mention	Web-based application
Programming Language	Not mention	Python	Not mention
Content	Social media posts of users	Survey data on the effects of COVID-19 on students' social life, mental health, and education.	Financial news
Dataset	Twitter data	The dataset contains 1182 student perspectives on 19 different features that represent different aspects of the survey.	Pre-processed datasets from digital platforms, provided text-based dataset, gathered finance-related datasets
Technique /Algorithm	Natural Language Processing (NLP), Sentiment Analysis, Machine Learning Algorithm.	AdaBoost, Shapley Additive Explanations (SHAP)	SLCNN, SVM, AdaBoost
Author	Yadav	Hussna et al.	S. Alanzi, A. Khaliq,

2.5 Data Vectorization Techniques

Data vectorization techniques are essential for converting textual information into numerical representations that can be processed by machine learning models. These techniques enable the extraction of meaningful features from text, allowing models to derive insights and make accurate predictions.

Methods such as bag-of-words, TF-IDF, and word embeddings transform text into numerical vectors, capturing word frequencies, importance, and semantic relationships. The choice of technique depends on the nature of the data and the specific task, and understanding these approaches is crucial for effectively analyzing and modeling text data in machine learning applications There will three data vectorization techniques that will be discuss in this subtopic.

2.5.1 Positive/Negative Frequencies

The Positive/Negative Frequencies represent the counts of specific words in a collection of tweets classified as positive (+) or negative (-). Each word is analyzed for its frequency in the positive and negative tweets separately.

Based on the Figure 2.4, this is the figure of example of the process positive/negative frequencies:

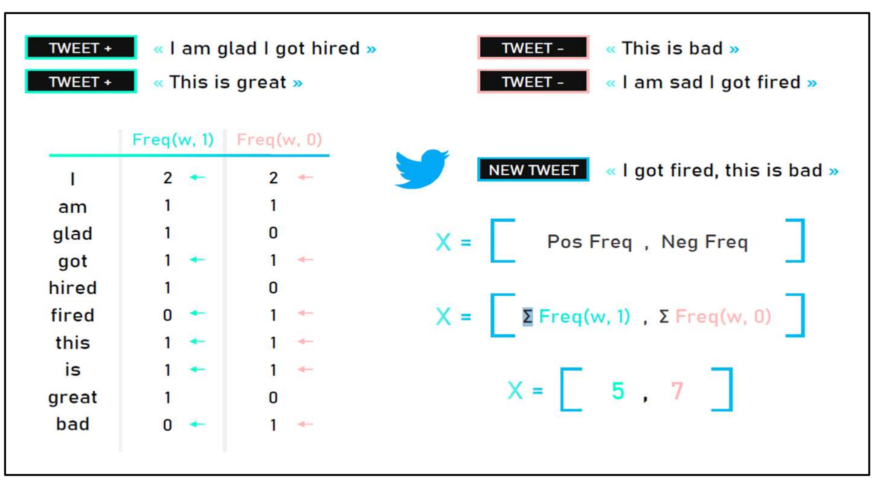


Figure 2. 4 Positive/Negative Frequencies
(Source: Termonia, 2021)

In Figure 2.4, it has four tweets which consists of two positive tweets and two negative tweets. The table in the Figure 2.4 shows the frequency of each word in the positive and negative tweets. For example, the word 'I' appears twice in both positive and negative tweets, so its frequency in the positive class (Freq(w,1)) is 2 and in the negative class (Freq(w,0)) is also 2(Termonia, 2021).

These frequencies provide insights into the occurrence of words in positive and negative contexts. Words that appear more frequently in the positive class than the negative class can indicate positive sentiment, while words that appear more frequently in the negative class suggest negative sentiment(Termonia, 2021).

To determine the overall positive and negative frequencies for a new tweet, it sum up the individual frequencies of the words present in that tweet. In the given example, the new tweet "I got fired, this is bad" has a positive frequency (Pos freq) of 5 (sum of the positive frequencies of the words 'I', 'got', and 'is') and a negative frequency (Neg freq) of 7 (sum of the negative frequencies of the words 'I', 'got', 'fired', 'this', and 'is'). These values can be used as features in a sentiment analysis algorithm to predict the sentiment of the new tweet(Termonia, 2021).

2.5.2 Bag-of-Words (BoW)

The Bag-of-Words (BoW) process is a common technique used in natural language processing to represent text data numerically. It involves creating a matrix representation of the text corpus, where each row corresponds to a tweet and each column represents a unique word found in the corpus(Termonia, 2021).

Figure 2.5 below show the picture of the BoW process:

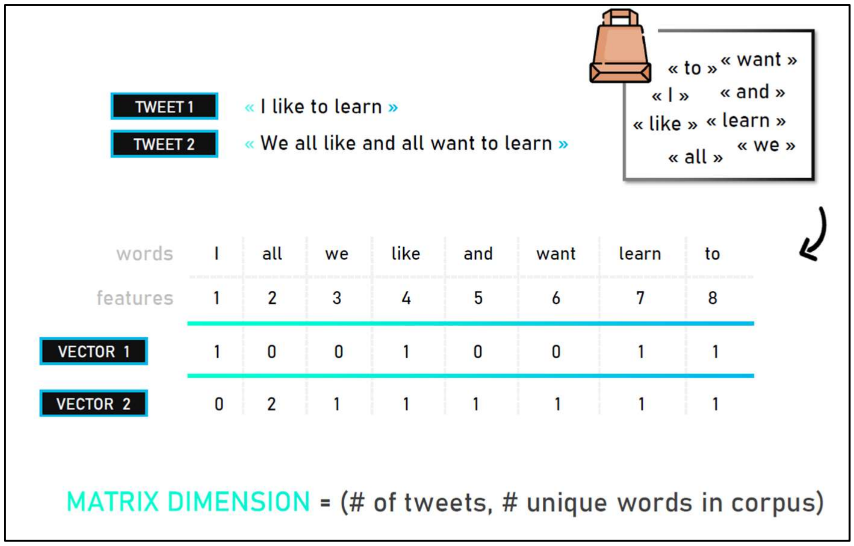


Figure 2. 5 Bag-of-Words (BoW)
(Source: Termonia, 2021)

Based on the Figure 2.5 above, it consists of two tweets: "I like to learn" and "We all like and all want to learn". The words from these tweets are used as features, and the matrix dimension is determined by the number of tweets and the number of unique words in the corpus(Termonia, 2021).

To construct the BoW matrix, we first identify all the unique words in the corpus. In this case, there are eight unique words: "I", "like", "to", "learn", "we", "all", "and", "want"(Termonia, 2021).

Each tweet is then represented as a vector where the values indicate the frequency or occurrence of the corresponding words in the tweet. For example, in the first tweet, the word "I" appears once, "like" appears once, and so on. Thus, the vector representation of the first tweet is [1, 1, 0, 1, 0, 0, 1, 1](Termonia, 2021).

By extending this process to all the tweets in the corpus, it then creates a matrix where each row represents a tweet, and each column represents a unique word.

The values in the matrix indicate the frequency or occurrence of the words in each tweet(Termonia, 2021).

The resulting BoW matrix provides a numerical representation of the text data that can be used as input for machine learning algorithms or other text analysis techniques. It allows us to analyze the presence and frequency of words across different tweets and enables us to derive insights and patterns from the text data(Termonia, 2021).

2.5.3 TF-IDF (Term Frequency-Inverse Document Frequency)

TF-IDF (Term Frequency-Inverse Document Frequency) is a method used to assess the significance of a term in a document within a collection of documents. It combines the concept of term frequency (TF), which measures how frequently a term appears in a document, and inverse document frequency (IDF), which measures the importance of a term in the entire document collection(Termonia, 2021).

TF is a measure of how frequently a term appears in a document. The formula for TF is as shown in the picture below, where "$n_{w,d}$" represents the number of occurrences of a term in a document, and "$\Sigma_k n_{w,d}$" represents the total number of occurrences of all terms in the document(Termonia, 2021).

TF assigns a higher value to terms that occur more frequently within a document, indicating their relative importance within that specific document(Termonia, 2021). The formula for the Term Frequency is as shown in the Figure 2.6 below.

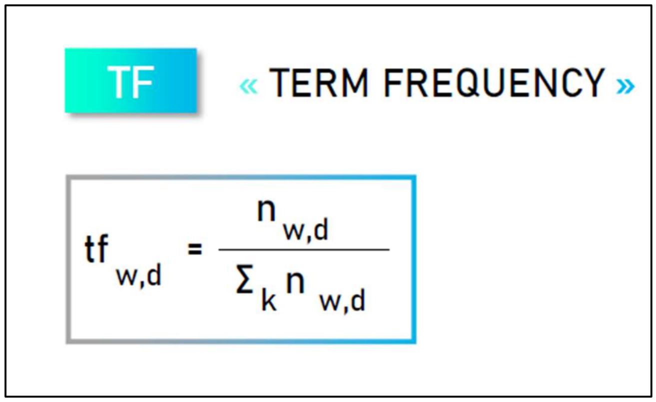


Figure 2. 6 Term Frequency Formula
(Source: Benjamin Termonia, 2021)

IDF measures the global importance of a term by calculating the logarithm of the ratio between the total number of documents in the collection and the number of documents that contain the term(Termonia, 2021).

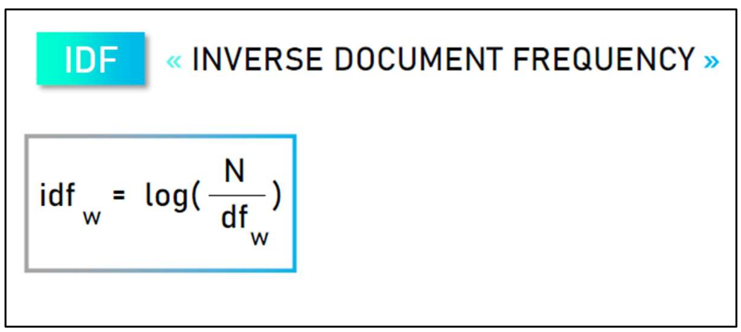


Figure 2. 7 IDF Formula
(Source: Termonia, 2021)

The formula for IDF is as shown in the Figure 2.7, where "N" represents the total number of documents in the collection, and "df_w" represents the number of documents that contain the term. IDF assigns a higher value to terms that appear in fewer documents, indicating their rarity and potential significance(Termonia, 2021).

2.6 Machine Learning Approaches

In the field of data modeling, choosing the most appropriate machine learning approach is crucial for achieving accurate and effective results. Machine learning approaches encompass a wide range of algorithms and techniques that enable computers to learn from data and make predictions or decisions without explicit programming. These approaches offer powerful tools for analyzing complex datasets, identifying patterns, and making informed decisions based on the available information.

2.6.1 Support Vector Machine

A Support Vector Machine (SVM) is a powerful machine learning algorithm used for classification and regression tasks. It is particularly effective in scenarios where the data points can be separated into different classes by finding an optimal hyperplane in a high-dimensional feature space. SVM aims to maximize the margin, or distance, between the hyperplane and the nearest data points of different classes, which helps in achieving better generalization and robustness(Medhat et al., 2017).

Its objective is to find a hyperplane in an N-dimensional space that effectively separates data points belonging to different classes. The key idea behind SVM is to maximize the margin, which is the distance between the hyperplane and the closest data points of each class(Rohith Gandhi, 2018). Based on the Figure 2.8 below, there are many possible hyperplanes that could be chosen to separate the classes of data points.

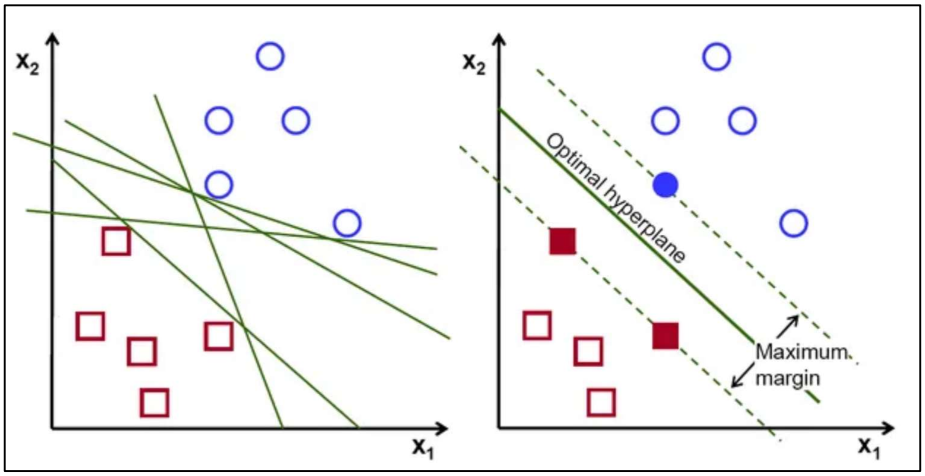


Figure 2. 8 Possible Hyperplane
(*Source:* Rohith Gandhi, 2018)

Hyperplanes are decision boundaries that separate data points, and the dimension of the hyperplane depends on the number of features in the dataset. For example, in a two-dimensional feature space, the hyperplane is a line, while in a three-dimensional space, it becomes a two-dimensional plane. The goal is to find the hyperplane with the maximum margin, as it provides a larger separation between the classes, leading to better generalization and improved performance on unseen data(Rohith Gandhi, 2018).

Figure 2.9 below shows the possible hyperplanes that has been separated in 2D and 3D feature space.

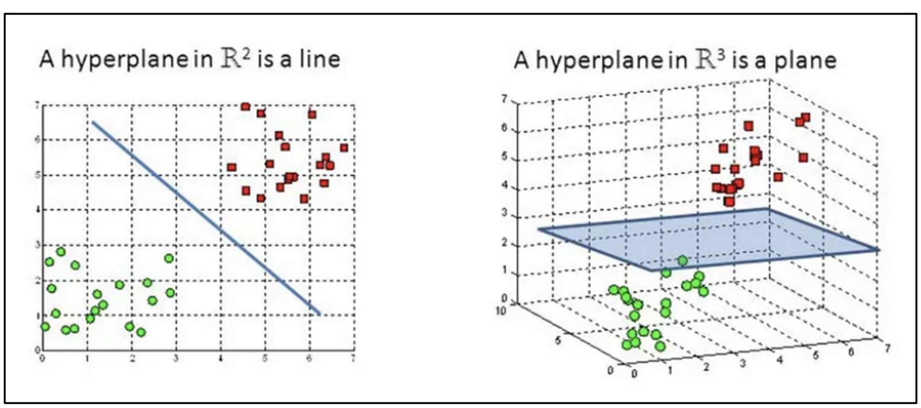


Figure 2. 9 Hyperplanes in 2D and 3D feature space
(*Source:* Rohith Gandhi, 2018)

Figure 2.10 below shows the support vector. Support vectors are data points that lie closest to the hyperplane and have the most influence on its position and orientation. These support vectors are crucial for defining the hyperplane and maximizing the margin. Removing or modifying support vectors can change the position of the hyperplane, impact the classification results and also helps in build the SVM(Rohith Gandhi, 2018).

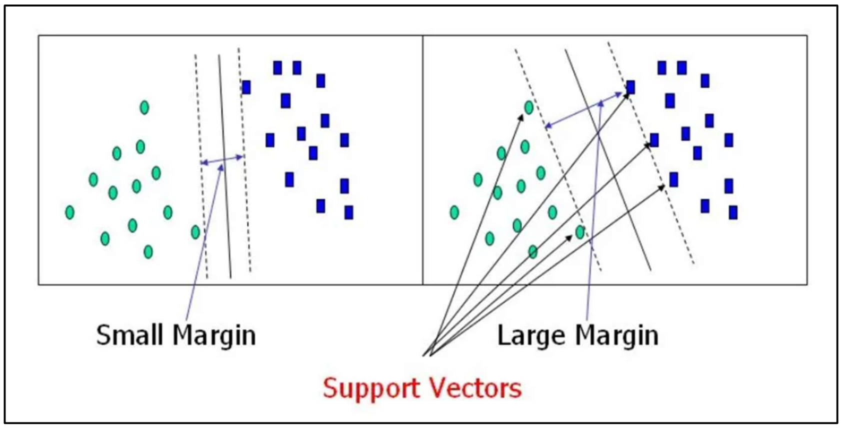


Figure 2. 10 Support Vectors
(*Source:* Rohith Gandhi, 2018)

LinearSVC is a variant of SVM that uses a linear kernel for classification. The linear kernel assumes that the data can be separated by a straight line (or a hyperplane in higher dimensions) in the feature space. LinearSVC finds the optimal hyperplane that best separates the data points while minimizing the classification errors(Medhat et al., 2017).

The LinearSVC model in scikit-learn is efficient for large-scale datasets and can handle high-dimensional feature spaces. It learns a linear decision boundary by solving an optimization problem that aims to maximize the margin between the classes, while also considering the misclassified samples. The model finds a set of support vectors, which are the data points closest to the decision boundary, and uses them to make predictions for new, unseen data points(Medhat et al., 2017).

SVM with a LinearSVC model is an effective approach for classification tasks that involves finding an optimal linear decision boundary to separate data points into different classes. It is a widely used algorithm due to its ability to handle complex datasets, handle large-scale problems efficiently, and its good generalization properties(Medhat et al., 2017).

In the SVM algorithm, the objective is to maximize the margin between the data points and the hyperplane. To achieve this, the loss function that will be use is hinge loss using the formula in Figure 2.11 below.(Rohith Gandhi, 2018). From the Figure 2.11 below, the function on the left can also be represented as a function on the right.

$$c(x, y, f(x)) = \begin{cases} 0, & \text{if } y * f(x) \geq 1 \\ 1 - y * f(x), & \text{else} \end{cases} \qquad c(x, y, f(x)) = (1 - y * f(x))_+$$

Figure 2. 11 Hinge loss function
(*Source:* Rohith Gandhi, 2018)

The cost associated with a prediction is 0 when the predicted value and the actual value have the same sign. If they differ, the loss value is calculated. Additionally, a regularization parameter is introduced in the cost function to balance the margin maximization and loss(Rohith Gandhi, 2018).

$$min_w \lambda \parallel w \parallel^2 + \sum_{i=1}^{n} (1 - y_i \langle x_i, w \rangle)_+$$

Figure 2. 12 Loss function for SVM
(*Source:* Rohith Gandhi, 2018)

After the loss function is have by using the formula in Figure 2.12, the partial derivatives are taken with respect to the weights to find the gradients. Using the gradients formula in Figure 2.13, it can update the weights(Rohith Gandhi, 2018).

$$\frac{\delta}{\delta w_k}\lambda \parallel w \parallel^2 \quad = 2\lambda w_k$$

$$\frac{\delta}{\delta w_k}\left(1 - y_i\langle x_i, w\rangle\right)_+ \quad = \begin{cases} 0, & \text{if } y_i\langle x_i, w\rangle \geq 1 \\ -y_i x_{ik}, & \text{else} \end{cases}$$

Figure 2. 13 Gradients

(*Source:* Rohith Gandhi, 2018)

When there is no misclassification, i.e. the model correctly predicts the class of the data point, it only have to update the gradient from the regularization parameter(Rohith Gandhi, 2018). Figure 2.14 shows the formula for the gradient update after there is no misclassification.

$$w = w - \alpha \cdot (2\lambda w)$$

Figure 2. 14 Gradient Update — No misclassification

(*Source:* Rohith Gandhi, 2018)

When there is a misclassification, i.e. the model make a mistake on the prediction of the class of our data point, it include the loss along with the regularization parameter to perform gradient update(Rohith Gandhi, 2018). Figure 2.15 below shows the gradient update after there is a misclassification.

$$w = w + \alpha \cdot (y_i \cdot x_i - 2\lambda w)$$

Figure 2. 15 Gradient Update — Misclassification
(*Source:* Rohith Gandhi, 2018)

2.6.2 K-Nearest Neighbour

The K-Nearest Neighbour (KNN) algorithm is a supervised learning method that can be used for classification or regression tasks. Its concept is simple: given a new data point, the algorithm identifies the k closest data points in the training set and assigns the label of the majority class among these neighbours to the new point. However, when data set is very large, the KNN algorithm may become computationally expensive and inefficient(Shokrzade et al., 2021).

To address this problem, researchers have developed parallized KNN algorithm that can be run on multiple processors or machines in parallel. These algorithms are often implemented using the distributed computing frameworks such as Hadoop, Spark, GPU. They are able to handle large data sets by dividing the data into smaller subsets and processing them simultaneously(Shokrzade et al., 2021).

But there are other difficulties with these parallelized KNN algorithms. For example, they require specialized hardware and may not be easily integrated into real-world or online applications. Additionally, they may be sensitive to the number of neighbours or the dimensionality of the data set, and the parameters used in these algorithms need to be optimized for best performance(Shokrzade et al., 2021).

Despite these challenges, parallelized KNN algorithms remain a promising approach for handling big data in classification and regression tasks. They offer a scalable and efficient solution for processing large data sets in parallel and can achieve high accuracy with appropriate tuning of their parameters. As such, they continue to be an important area of research in machine learning and data science(Shokrzade et al., 2021).

The Figure 2.16 below shows the example of creating the KNN finding tree.

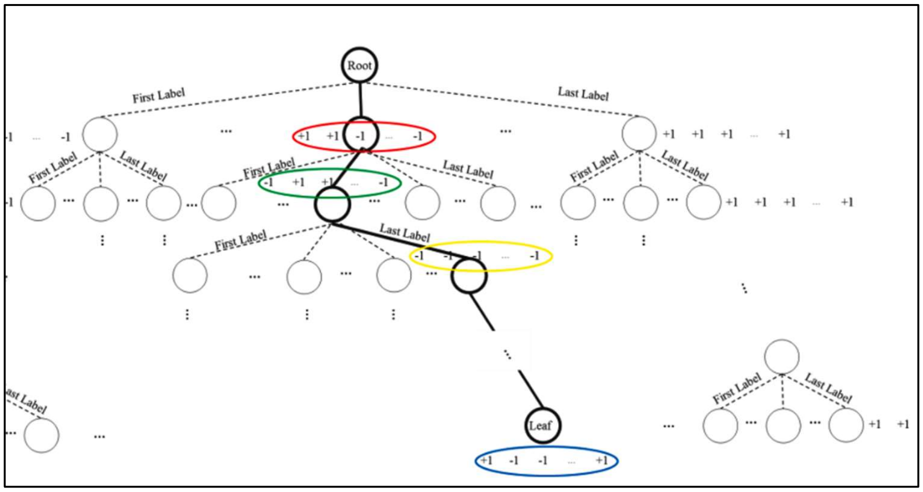


Figure 2. 16 Creating the KNN figure tree
(*Source*: Shokrzade et al., 2021)

2.6.3 Logistic Regression

In many disciplines, including engineering, medicine, marketing, and more, logistic regression (LR) is a sort of generalised linear regression analysis model. Logistic Regression can be effective for sentiment analysis, but its effectiveness depends on various factors such as the quality and size of the dataset, feature selection, and hyperparameter tuning. Logistic Regression is a linear model that is simple and interpretable, making it a popular choice for sentiment analysis tasks(Guo, 2022).

It is used to predict the likelihood of a situation where there are only two possible outcomes (dichotomy), as well as to clarify the relationship between a dependent variable and one or more independent variables at the nominal, sequential, interval, or ratio level. The history of the LR model's evolution is comprehensive, and it has been thoroughly evaluated by researchers, who have also studied it during the past 15 years(Guo, 2022).
Figure 2.17 below represent logistic regression in a graph.

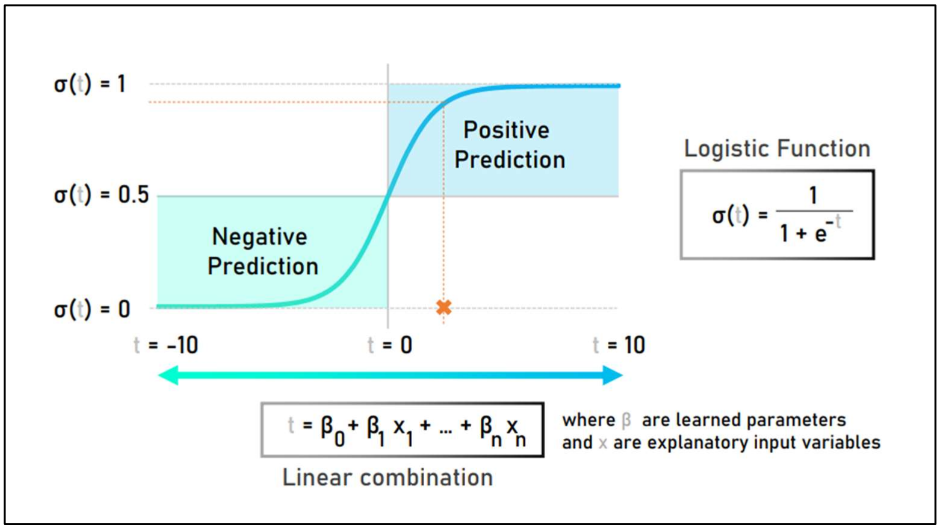


Figure 2. 17 Logistic Regression Graph
(*Source: Benjamin Termonia, 2021)*

In logistic regression, the main objective is to predict the probability of an event happening, for example like a person having a mental health problem, based on one or more input variables (e.g., tweets containing certain keywords). To do this, logistic function is used, which has an S-shaped curve (Benjamin Termonia, 2021).

The equation for the logistic function is σ(t) = 1 / (1 + e^(-t)), where t is the linear combination of the input variables and the learned parameters. This means that the output of the logistic function is always between 0 and 1, representing the probability of the event happening (Benjamin Termonia, 2021).

In a logistic regression graph, it can plot the probability of the event happening (y-axis) against the input variable(s) (x-axis). The curve of the logistic function will start at 0 when the input variable(s) are very low and end at 1 when the input variable(s) are very high. The curve will be steepest when the probability is close to 0.5, meaning that a small change in the input variable(s) can have a big impact on the predicted probability (Benjamin Termonia, 2021).

The goal of logistic regression is to learn the optimal values of the parameters β so that the predicted probabilities match the actual outcomes in the training data. This allows us to use the learned model to predict the probability of the event happening for new data points (Benjamin Termonia, 2021).

2.6.4 Summarization of the Machine Learning.

This section provides a comprehensive summary of how machine learning techniques are applied in sentiment analysis, with a focus on three specific methods: K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Logistic Regression. Sentiment analysis is a valuable tool that helps us understand people's feelings and opinions expressed in text data. By examining the advantages, disadvantages, and suitability of each machine learning approach for sentiment analysis, this project aims to gain a clear understanding of how these methods can be effectively used to extract valuable insights from text and determine sentiments towards various subjects.

The summarization of the machine learning approaches can be seen in the Table 2.2 on the next page.

Table 2. 2 Comparison between KNN, SVM and Logistic Regression

Algorithm	**Advantages**	**Disadvantages**	**Suitability for Sentiment Analysis**
K-Nearest Neighbours (KNN)	Easy to implement and understand	Sensitive to noise and outliers	May not be suitable for sentiment analysis because of the high dimensionality of text data
Support Vector Machine (SVM)	Effective for handling high-dimensional data and text classification	Can be computationally intensive and slow to train	Suitable for sentiment analysis because of its ability to effectively model non-linear relationships between the input features and the binary sentiment labels
Logistic Regression	Simple and easy to interpret model	Not suitable for complex relationship between variables	Suitable because it can effectively model the relationship between the input features and the binary sentiment labels

Based on the above comparison, Support Vector Machines (SVM) with LinearSVC is chosen for depression prediction due to its effectiveness in text classification and its ability to handle high-dimensional data. SVMs are known for their robustness to noise and their capability to handle large-scale datasets. In sentiment analysis, where text data can be noisy and high-dimensional, LinearSVC excels at learning the relationships between features and labels.

2.7 Platform Development

This section is dedicated to the comprehensive exploration of platform development, focusing on the creation of a web-based system. It covers three essential aspects of the development process: the frontend of the web platform, the backend of the web platform, and analysis of the overall system platform.

2.7.1 Frontend of a Web Platform

The frontend of a web platform is the part of the system that interacts directly with users. It is responsible for displaying content and user interface elements such as forms, buttons, and menus, and for handling user interactions such as clicks and inputs. The frontend typically runs in a user's web browser and communicates with the backend through API calls(GmbH, 2020).

In terms of the specific technologies, HTML and CSS are the main building blocks of the frontend. HTML provides the structure and content of web pages, while CSS is used to style and layout those pages. Together, they determine the visual appearance and layout of the user interface(GmbH, 2020).

The frontend will be responsible for presenting the user interface and interacting with the user. It will provide a user-friendly interface for users to interact with your web application. The frontend will include components such as forms, buttons, input fields, and other user interface elements that allow users to input data and interact with the web application(GmbH, 2020).

The frontend will also provide visualization components to display the results in a clear and understandable way. For example, it might display a pie chart showing the percentage of positive and negative, sentiment in the analysed text(GmbH, 2020).
In summary, the frontend of the web platform is the part of the system that users interact with directly, and it is primarily built using HTML and CSS. Other tools such as Jupyter Notebook and Visual Studio Code may also be used

for certain frontend development tasks, depending on the specific requirements of your project.

2.7.2 Backend of a Web Platform

In the context of this project, the backend development will focus on performing sentiment analysis for depression detection using the provided dataset. The sentiment analysis is conducted using machine learning libraries such as Scikit-learn, pandas, and nltk.

The code provided implements a pipeline that includes data cleaning, data vectorization, and data modeling using a SVM with LinearSVC(Support Linear Classifier) classifier. The pipeline is optimized using GridSearchCV to find the best model and parameters for the task. The accuracy of the model is calculated to evaluate its performance.

Additionally, the code includes visualization of the model's performance metrics, such as the confusion matrix and classification report, using seaborn. These visualizations provide insights into the model's ability to correctly classify depression and non-depression texts.

Although the code does not explicitly store the data in a database, it performs sentiment analysis on the provided dataset to detect depression. The focus of the backend development is on data processing, modeling, and visualization to enable effective depression detection using sentiment analysis techniques.

The backend will also include the development of APIs that enable the frontend to access and retrieve data from the database. Additionally, sentiment analysis algorithms will be implemented on the server-side to perform sentiment analysis on social media data. This will involve using machine learning libraries and frameworks such as Scikit-learn libraries, pandas, and nltk.

To achieve these goals, several software tools will be used. The choice of software tools will depend on the specific needs of the project and the programming languages and frameworks selected. The backend development tools that could be used in this project include Flask as web frameworks, and Python and its libraries for sentiment analysis.

2.7.3 Analysis on System Platform

A system platform is a software framework that provides a set of common services and functionalities that can be used to develop and deploy applications. It typically includes components such as an operating system, programming languages, libraries, and tools for developing, testing, and deploying software applications (Platforms vs. Applications: What's the Difference?, 2023).

In comparison, a mobile platform is a type of system platform that is designed specifically for developing applications that can run on mobile devices such as smartphones and tablets. It typically includes an operating system, programming languages, libraries, and tools that are optimized for mobile devices (Platforms vs. Applications: What's the Difference?, 2023).

When it comes to using a web-based system, the main difference between a system platform and a mobile platform is the level of optimization for mobile devices. Since a web-based system is accessed through a web browser, it can run on any device that has a browser installed, regardless of the operating system or hardware platform (Consequences of Student Mental Health Issues, 2020).

As a result, the system platform used for a web-based system may not be optimized for mobile devices in the same way that a mobile platform is. However, many system platforms and frameworks, such as React, Angular, and Vue, offer features and tools that can help developers create responsive and mobile-friendly web applications (Platforms vs. Applications: What's the Difference?, 2023).

2.8 Chapter Summary

This literature review section discusses about the overview of the mental health disorder among university student, machine learning approaches, related works and the platform development. As a result, SVM (Support Vector Machines) is chosen for this project based on this literature review due to its effectiveness in binary classification tasks, such as determining whether a user is exhibiting symptoms of depression based on their input text.

SVM is known for its ability to handle high-dimensional data and separate classes with a clear margin, making it suitable for sentiment analysis. It offers good generalization capabilities and performs well even with a limited amount of training data.

Other than that, SVM is less prone to overfitting and can handle large feature spaces efficiently. Overall, SVM's strong classification performance and its ability to handle complex data make it a suitable choice for this project's sentiment analysis task. For the analysis on system platform, it has been decided that this project will be focusing on web based.

CHAPTER 3

METHODOLOGY

The research methodology that was used to develop and design this project were covered in this chapter. The system development process, from the requirements phase to the deployment and review phase, is thoroughly described. Explanations and justifications for the research design and data collection methodologies were given to justify the approach. Moreover, this chapter includes the hardware, software and data requirements that is essential to accomplishing the project's system development goals.

3.1 Introduction

Software development methodology refers to the approach or framework that a software development team follows to develop a software product. A software development methodology can aid in streamlining the process, cutting down on the time and effort needed to produce software. It can also help to ensure that the process is consistent and repeatable, which can improve the quality of the finished result(Chaudhari & Joshi, 2021).

A method called the Software Development Life Cycle (SDLC) explains the stages that occur while developing a software product. It includes planning, analysis, design, implementation, testing, deployment, and maintenance. Different software development methodologies will have their own variations of the SLDC, but the overall process is the similar. The goal of the SLDC is to produce high-quality software that meets the needs of the end users and is delivered on time and withing the budget(Chaudhari & Joshi, 2021).

3.2 Agile Methodology

One of the most popular frameworks for software development processes is agile methodology. Agile methodology is often chosen for projects that require flexibility and adaptability. In my mental health website project, some changes may be encountered in requirements or new features that need to be added. Agile methodology can help to respond to these changes quickly and efficiently(Chaudhari & Joshi, 2021).

Most frequently, the agile technique handles the frequently changing requirements in the software project. Modern Agile methodologies always prioritise quick software delivery and adaptation to constantly changing requirements. It also features systematic releases and small iterations. It will improve software quality, effectiveness, and throughput (Chaudhari & Joshi, 2021).

The phases of the agile model are shown in Figure 3.1.

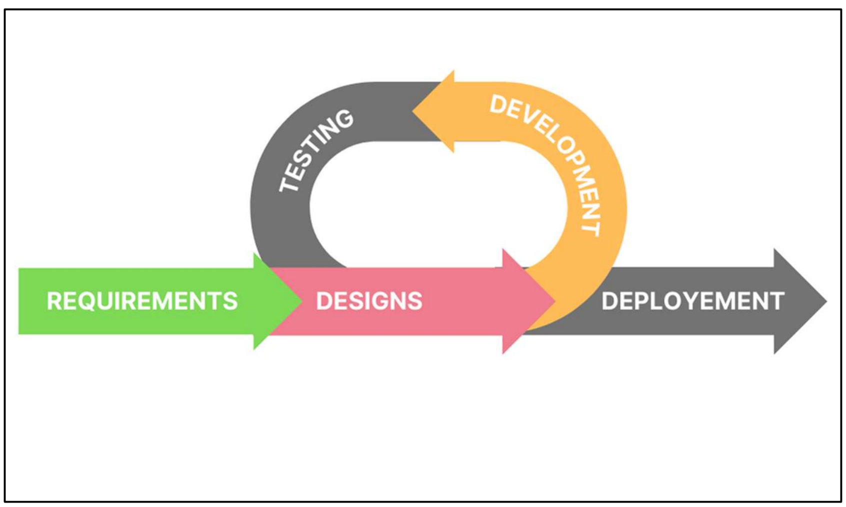


Figure 3. 1 Agile Model
(*Source*: Salza et al., 2019)

The Agile methodology is based on iterative and incremental development, which involves breaking down the project into smaller, more manageable phases. The phases of Agile methodology include requirements, design, development, testing, and deployment. In the requirements phase, the focus is on gathering the requirements. This involves understanding the needs of the users and defining the features that will be included in the project. In the design phase, the focus is on creating a high-level design that describes how the project will meet the user's needs. This phase includes creating wireframes, mock-ups, and prototypes.

In the development phase, the focus is on building the product in small, incremental pieces. This allows the me to continuously test and validate the product as it is being built. In the testing phase, the focus is on validating that the project meets the user's needs and is free of defects. Finally, in the deployment phase, the focus is on releasing the project to the users.

For a mental health website project, Agile methodology is an appropriate choice as it allows for a flexible and iterative approach to development. The requirements phase is crucial in this project as it involves understanding the needs of the users, which can be complex in the case of mental health. The design phase is important in ensuring that the website is user-friendly, accessible, and meets the needs of the users.

The development phase allows for continuous testing and validation of the website, which is essential in ensuring that it is safe, reliable, and effective in supporting the mental health of users. Finally, the testing and deployment phases allow for the website to be tested and released to users in a controlled and effective manner.

Deliverables for this phase might include a detailed plan for collecting and pre-processing data from Twitter and a list of key topics and issues related to mental health that will be addressed on the website system.

3.2 Research Framework Overview

In this section, it represents the comprehensive summary of the research framework for the development of a sentiment analysis system using the Agile methodology. The primary objective of this research is to create an efficient and user-friendly system that can analyze sentiments expressed in textual data, providing valuable insights into people's feelings and opinions. The research framework encompasses various phases, including requirement analysis, design, implementation, testing, and deployment. Throughout this summary, it aims to highlight the significance of each phase in ensuring the successful development and deployment of the sentiment analysis system, contributing to the broader understanding of sentiment analysis techniques and their practical applications.

Table 3.1 below shows the summary of research framework that will be conducted using the agile model.

Table 3. 1 Summary of Research Framework using Agile Methodology

Phase	**Activities**	**Deliverable**
Requirement Analysis	▪ Identify the area and the purpose of the proposed system implementation. ▪ Set the project title. ▪ Define the problem statements and significance of this project. ▪ Identify the objectives and scope	▪ Background of Study ▪ Problem Statements ▪ Project Objectives ▪ Project Scope ▪ Project Significance
	▪ Collect all relevant information for the project. ▪ Review existing journal, article, and book. ▪ Review the existing related projects	▪ Literature Review

Design	▪ Identify project methodology. ▪ Define the technique and procedure for the development of the proposed system. ▪ Design flowchart, use case and user interface	▪ Agile Methodology ▪ Flowchart ▪ Use Case ▪ User Interface Design
Development	▪ Develop the model for the Sentiment Analysis using Support Vector Machine (SVM) ▪ Build a system prototype. ▪ Define the hardware and software requirement	▪ Hardware Requirements ▪ Software Requirements ▪ Data Requirements
Testing	▪ Test all available functions to ensure all requirements are met and work well together. ▪ Conduct system acceptance testing, including the functionality and usability test. ▪ Log any error or glitch found in the system	▪ Functionality Testing ▪ Usability Testing ▪ Accuracy Testing
Deployment and Review	▪ Prepare the system for deployment, ensuring all necessary files and configurations are in place. ▪ Evaluate the systems strengths, limitations of the projects and provide the future recommendations.	▪ Deployed System ▪ Strengths of the project ▪ Limitations ▪ Future Recommendations

3.3 Requirements Analysis Phase

The Requirement Analysis phase is an essential initial step in developing the proposed system. During this phase, a comprehensive understanding of the project's purpose and objectives is sought. The specific area of system implementation is identified, and its purpose, along with the significance of addressing mental health support, is defined. Additionally, a clear project title is set to indicate the system's focus.

To establish a well-defined project scope, the objectives to be achieved and the boundaries within which the system will operate are identified. This includes providing background information, stating specific problem statements that the system intends to tackle, and defining the project's objectives and scope.

Collecting all relevant information for the project is imperative as it forms the basis for system development. This involves conducting a thorough literature review, exploring existing journals, articles, and books related to sentiment analysis and depression detection. Analyzing prior research and examining related projects provides valuable insights and identifies potential gaps in the field, which will inform the system's approach and design.

3.4 Design Phase

In the Design Phase, the focus is on shaping the framework and structure of the proposed system. Firstly, the project methodology is identified, with the Agile Methodology chosen for this project. The Agile approach emphasizes iterative development and collaboration, allowing for adaptability to changing requirements and the delivery of a user-centric solution effectively.

Next, the technique and procedure for system development are defined. This involves outlining the steps, methodologies, and best practices to be followed during implementation, ensuring a systematic and organized approach.

Another critical aspect of the Design Phase is creating essential design elements, including flowcharts, use cases, and user interface (UI) design. Flowcharts visually represent the flow of processes and interactions within the system, providing a clear overview of its functionality. Use cases outline various scenarios and interactions between users and the system, helping to understand user requirements and system behaviour effectively. Lastly, the user interface design focuses on creating a visually appealing and intuitive interface that ensures a seamless user experience.

By addressing these aspects in the Design Phase, a strong foundation is laid for the successful implementation of the proposed system. The Agile Methodology allows for flexibility and collaboration, while the defined techniques and procedures ensure a well-structured and systematic development approach. The design elements, such as flowcharts, use cases, and user interface design, contribute to the system's effectiveness, usability, and overall user satisfaction.

3.5 Development Phases

In the Development Phase, the primary focus is on creating an effective Sentiment Analysis model specifically designed to detect depression in text data. Sentiment analysis is a powerful technique that allows us to gain insights into people's emotions and opinions expressed through written text. For this project, it aims to harness this technique to detect whether a given text exhibits symptoms of depression or not. To achieve this, a robust model will be develop using the Support Vector Machine (SVM) with LinearSVC algorithm, which is known for its effectiveness in binary classification tasks like ours.

By diligently addressing these development aspects, this project aims to create a robust and efficient solution that can detect depression symptoms effectively from text data. The ultimate goal is to contribute to the field of mental health support by providing a valuable tool that aids in identifying potential signs of depression, enabling timely intervention and support for those in need.

3.6 Testing Phases

In the Testing Phase, it focuses on ensuring the reliability and effectiveness of the developed system, specifically the sentiment analysis model for detecting depression in text. This phase plays a crucial role in validating the system's functionalities and overall performance, making sure it meets all the specified requirements and functions seamlessly.

To begin, first the functionality testing will be conduct, which involves a comprehensive examination of all available functions in the system. By doing so, it verifies that each function operates as intended and that all the requirements set during the development phase are successfully met. This ensures that the system performs its core tasks accurately and consistently.

Next, proceed with the usability testing, aiming to assess the system's user-friendliness and accessibility. This involves having potential users interact with the system to evaluate how intuitive and easy it is to navigate and utilize. By gathering user feedback, any potential areas can be identified for improvement in the system's design and user experience.

Additionally, accuracy testing will be performed to evaluate the model's performance in detecting depression symptoms in text data. This involves testing the system with a diverse dataset to measure its precision and recall in correctly identifying positive and negative instances of depression. Accuracy testing is crucial as it determines the system's reliability in real-world scenarios, ensuring that it provides meaningful results with high precision.

Throughout the testing process, it meticulously log any errors, glitches, or unexpected behaviour found in the system. The comprehensive logging helps in tracking and addressing issues effectively, making necessary adjustments and improvements to enhance the system's overall performance and stability.

By conducting functionality, usability, and accuracy testing, it ensures that this developed system is robust, user-friendly, and capable of effectively detecting depression symptoms in text data. The thorough testing process helps validate the system's readiness for deployment, ensuring that it can make a meaningful contribution to the field of mental health support by identifying potential signs of depression and enabling timely intervention and support for individuals in need.

3.7 Deployment and Conclusion Phases

In the Deployment and Review phase, it focuses on preparing the system for actual deployment by ensuring all necessary files and configurations are in place. This marks a crucial transition from the development environment to the live environment where users will access the system.

After deployment, moves to evaluating the project's strengths and limitations. This involves carefully assessing the system's successes and accomplishments, such as its accuracy in detecting depression symptoms and user-friendliness. We also identify areas where the system may encounter challenges or fall short to understand its limitations better.

Based on the evaluation, it provides future recommendations to enhance the system. These recommendations aim to guide continuous improvement, expansion, and adaptation to better meet users' needs and address any identified limitations.

3.8 Project Timeline

A project timeline is a visual representation of the planned start and end dates of the various tasks or activities that make up a project. It may include information about the resources that are assigned to each task, the dependencies between tasks, and any milestones or deliverables. In creating a project timeline, it is important to consider the scope of the project, the resources that are available, and any potential risks or challenges. The timeline for this project has been shown in the Gantt charts that is provided in the Appendices section in Appendix A.

3.9 Chapter Summary

This chapter concludes by summarising the project's methodology, which led to the project's successful completion. The phases of the agile model technique are chosen because they work well for creating web applications. The five phases of the agile methodology are requirements, design, development, testing, and deployment.

This chapter goes into the detail about each process. Making sure the system produces the desired results at the end depends on each stage. Regardless of the methodology chosen, it is important to carefully plan and execute the software development process to ensure that the final product meets the needs of the users and delivers value to the business. Ongoing evaluation and improvement of the methodology can also help ensure that it is effective and efficient.

CHAPTER 4

SYSTEM DESIGN AND DEVELOPMENT

In this chapter, it will focus on the crucial stages of system design and development for our mental health website. It begins by analysing the system requirements for this project, developing the website, implementing functionalities and features while ensuring a seamless user experience. Through these stages, this project aims to create a user-centric mental health website that provides a supportive and informative space for individuals seeking mental health resources and support.

4.1 Analysis of the System Requirements

The analysis requirements of a system encompass various aspects that need to be considered to ensure its successful development and operation. This includes the examination of hardware requirements, software requirements, and data requirements.

By thoroughly analysing these requirements, it can ensure that the system is built on a solid foundation, with the necessary hardware, software, and data components in place to meet its objectives and deliver the desired functionality and performance.

4.1.1 Software Requirements

This section will discuss about the software requirements for the mental health website system. These software components and tools are essential for analyzing and designing the mental health website system, facilitating tasks such as coding, responsive design, document creation, and system modeling. Table 4.1 below shows the list of software tools that has been used during the system development.

Table 4. 1 Software Requirements

Software	Description
Google Colab	Google Colab is a cloud-based development environment that provides free access to computational resources and allows collaborative coding. It is ideal for running and developing machine learning models and data analysis tasks
Mobirise CSS	Mobirise CSS is a responsive web design framework that enables the creation of mobile-friendly websites. It provides a set of CSS classes and components that adapt to different screen sizes, ensuring a seamless user experience on various devices.
PyCharm	PyCharm is an integrated development environment (IDE) specifically designed for Python programming. It offers advanced coding features, debugging tools, and project management capabilities, making it efficient for developing Python-based applications.
Visual Studio Code	Visual Studio Code is a comprehensive IDE developed by Microsoft. It supports multiple programming languages and provides a rich set of tools for building and debugging applications. It offers a wide range of features, including code editing, version control, and project management.
Microsoft Word	Microsoft Word is a popular word processing software that allows the creation and editing of documents. It is commonly used for writing reports, documentation, and other textual content.
StarUML	StarUML is a modeling tool used for creating UML (Unified Modeling Language) diagrams. It provides a graphical representation of the system's structure and behaviour, facilitating the analysis and design of software systems.

4.1.2 Hardware Requirements

The hardware requirements for a software system are essential to ensure smooth operation and optimal performance. In this project, we need to consider the hardware components that will support the system's functionalities. This includes the laptop model, processor type, operating system, RAM capacity, HDD and SSD storage, and system architecture. Table 4.2 below shows the hardware requirements for this project.

Table 4. 2 Hardware Requirements

Hardware	**Description**
Laptop Model	HP Laptop - 15s-eq1149au
Processor	AMD Ryzen 5 4500U with Radeon Graphics @ 2.38GHz
Operating System	Windows 11
Memory (RAM)	8.00 GB
Hard Disk Drive (HDD)	500 GB
Solid State Drive (SSD)	512 GB
System Architecture	64-bit

4.1.3 Data Requirements

The data requirements for this project involve collecting tweets using the Twitter API. The collected data consists of random tweet texts from Twitter users. These tweets are labelled as non-depressive text with a label of 0.

For the depressive text, the data from studies conducted by researchers Uddin et al., (2022) has been obtained. According to this researcher, the dataset has been approved by professional experts such as doctors, psychologists, and nurses, who have determined that the text exhibits symptoms of depression. These data sources are labelled as depressive text with a label of 1 and will be used to train and test the sentiment analysis model for this project, same goes with data sources that was labelled as non-depressive text.

From the Figure 4.1 below, it shows the sample dataset for the text that contain depressive text or exhibit depression symptoms:

Text	Sentiment
I'm losing my will to live.	1
The fear of asking for extensions or accommodations due to the stigma attached to mental health issues. #StigmaInAcademia	1
The fear of failure and disappointing others, leading to paralyzing academic anxiety. #FearOfFailure	1
I can't escape the grip of this deep sadness.	1
"Depression has affected my self-esteem and made me believe I'm not worthy of love or happiness." #SelfEsteemStruggles	1

Figure 4. 1 Sample dataset for depressive text

Figure 4.2 below shows the sample dataset for the text that does not contain any depression symptoms:

Text	Sentiment
"I had a fantastic day at a food festival, savoring a variety of cuisines, sampling gourmet dishes, and experiencing culinary delights."	0
"Your dreams are worth pursuing. Stay committed, work hard, and never give up."	0
"The universe has a way of aligning things in divine timing. Trust the process and have faith in the journey."	0
@hi_sweetye I hope so	0
ch day is an opportunity to create a life that aligns with my values, passions, and desires. I have the power to shape my own destiny."	0

Figure 4. 2 Sample dataset for non-depressive text

4.2 System Design

The Design Phase is a pivotal stage in the development for this project, marking a critical step towards turning the ideas into reality. In this phase, it establishes the framework for the system's efficient and effective implementation, shaping its structure and functionalities. Additionally, the technique and procedure are defined so that it will guide the system's development, utilizing flowcharts to visualize the overall flow of the system. By creating detailed use cases it can outline the system's functionalities, while designing user interfaces that prioritize user experience and accessibility.

4.2.1 Use Case Diagram

A use case diagram is a type of Unified Modelling Language (UML) diagram that shows the relationships between actors and the actions they can perform within a system.

It is used to represent the requirements of a system and to visualize the different scenarios that the system must be able to handle. In a use case diagram, actors are represented by stick figures, and the actions they can perform are represented by ovals.

Figure 4.3 below is the use case diagram for the mental health website:

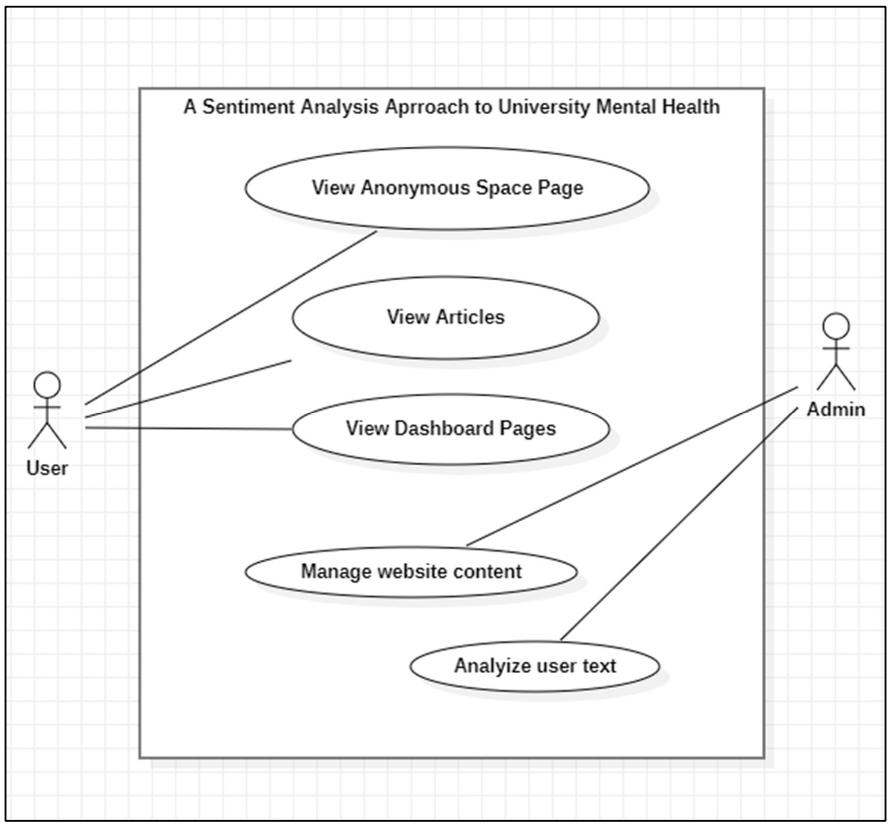


Figure 4. 3 Use Case Diagram

Actors:

a) User: a person who visits the website and uses its features

b) Admin: a person who manages the website and its content

Use cases:

a) View Mental Health Resources Page (by User)

This use case describes the action of a user accessing the Mental Health Resources page on the website. The user can browse through various resources educational content related to mental health. This use case aims to provide users with valuable information and support for their mental well-being.

b) Anonymous Space Page/Sentiment Analysis Page (by User)

In this use case, the user interacts with the Anonymous Space or Sentiment Analysis page. The user can anonymously share their thoughts, feelings, or experiences, and the system provides sentiment analysis on their input text. This use case aims to create a safe and supportive environment for users to express themselves and receive insights based on sentiment analysis.

c) View Dashboard Pages (by User):

In this use case, the user interacts with the Dashboard Pages on the website. The user can access visualizations and analysis derived from the dataset that powers the depression prediction using sentiment analysis. The Dashboard Pages present various visual representations such as pie charts, bar charts, and word clouds. These visualizations showcase sentiment distributions, common words in the dataset, and distinct patterns between depression and non-depression categories. This use case aims to provide users with a deeper understanding of the sentiment analysis results and insights from the dataset in a visually appealing and informative manner. By exploring the Dashboard Pages, users can gain valuable insights into the sentiment patterns and characteristics associated with depression prediction.

d) Manage Website Content (by Admin):

This use case involves the administrative role of managing the website's content. The admin can add, update, or remove articles, personal stories, self-assessment tools, and other resources. They can also manage user-generated content, moderate discussions, and ensure the accuracy and relevance of the information provided on the website. This use case allows the admin to maintain the quality and integrity of the website's content.

e) Analyze User Text (by Admin):

In this use case, the admin analyzes the user's input text from the Anonymous Space or Sentiment Analysis page. The admin utilizes sentiment analysis algorithms or tools to process the text and extract sentiment-related information. This analysis can provide valuable insights into user sentiments, and potential areas of concern.

These use cases illustrate the different interactions and functionalities of the mental health website, catering to both user engagement and administrative tasks. By addressing these use cases, the website aims to provide a comprehensive and user-friendly platform for mental health support and resources.

4.2.2 Flowchart of the Sentiment Analysis

The sentiment analysis flow is a crucial component of the mental health website system. It involves processing and analyzing textual data obtained from various sources, such as the data from studies conducted by researchers Uddin et al., (2022) and Twitter API, to determine the sentiment expressed in the text. In this section, it will present the flowchart illustrating the different stages of the sentiment analysis process.

Figure 4.4 below shows the Sentiment Analysis process consists of several essential steps that enable this project to analyze sentiments expressed in textual data, particularly in the context of Twitter data. Let's go through each step:

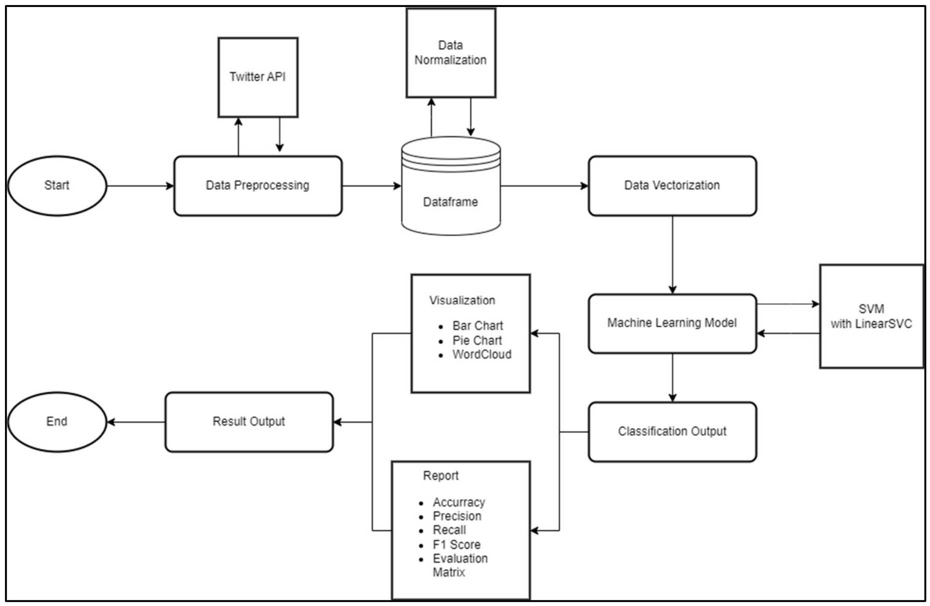


Figure 4. 4 Flowchart of the Sentiment Analysis Process

1. Data Pre-processing
   The first step in the sentiment analysis flow is data pre-processing, which involves preparing the raw tweet data for analysis. The data is obtained from the Twitter using its Application Programming Interface (API) for collecting random tweets that will be labelled as non-depressive texts and data from studies conducted by researchers Uddin et al., (2022), labelled as the depression texts. It consists of text and sentiment result, which will be used to train and evaluate our sentiment analysis algorithm.
2. Storing data in a Dataframe and do the Data Normalization/Cleaning
   After data collection, the tweets are stored in a dataframe, a tabular data structure suitable for analysis. Before proceeding, data normalization and cleaning are performed. This involves removing unnecessary characters, symbols, and special characters, and handling

misspellings or abbreviations to ensure consistent and clean data for analysis.

3. Data Vectorization

   To analyze textual data, it is necessary to convert the text into numerical vectors. Data vectorization is employed to achieve this. Two specific techniques used are TfidfVectorizer and SelectKBest. TfidfVectorizer calculates the importance of each word in a tweet based on its frequency and rarity in the dataset. SelectKBest is a feature selection technique that retains only the most relevant and informative words for analysis, enhancing the efficiency and accuracy of the model.

4. Machine Learning Model using SVM with LinearSVC

   The sentiment analysis process involves training a machine learning model using the Support Vector Machine (SVM) algorithm with LinearSVC (Linear Support Vector Classification). SVM is a powerful algorithm for binary classification tasks, such as sentiment analysis. By inputting the vectorized data into the SVM model, it learns to classify text into positive or negative sentiment categories.

5. Classification Output

   Once the SVM model is trained, it can be used to classify new and unseen data. The sentiment analysis process assigns sentiment labels to the input text, indicating whether the text expresses positive or negative sentiment.

6. Generating Visualization of the Data

   To gain insights into the analysis results, data visualizations such as bar charts, pie charts, and word clouds are generated. These visualizations provide a better understanding of the sentiment distribution in the data and the most frequently occurring words in different sentiment categories.

7. Reporting Evaluation Metrics

   To assess the model's performance, a comprehensive report is generated, including evaluation metrics such as accuracy, precision, recall, and F1 score. These metrics provide valuable information about the model's effectiveness in accurately classifying sentiments.

8. Result Output

   The sentiment analysis process produces the final output, including the sentiment labels assigned to each tweet and the corresponding evaluation metrics. This output helps understand the sentiment distribution in the data and the performance of the sentiment analysis model in classifying sentiments accurately.

4.2.3 Flowchart of the System

The flowchart of the system outlines the sequential steps and interactions within the mental health web platform with sentiment analysis. It provides a visual representation of how users navigate through the platform and interact with its various features. The flowchart showcases the process of anonymous text input, sentiment analysis, and the resulting prediction. The flowchart serves as a clear and intuitive guide to understanding the platform's functionalities and the overall user experience. Figure 4.5 below shows the flowchart of the system.

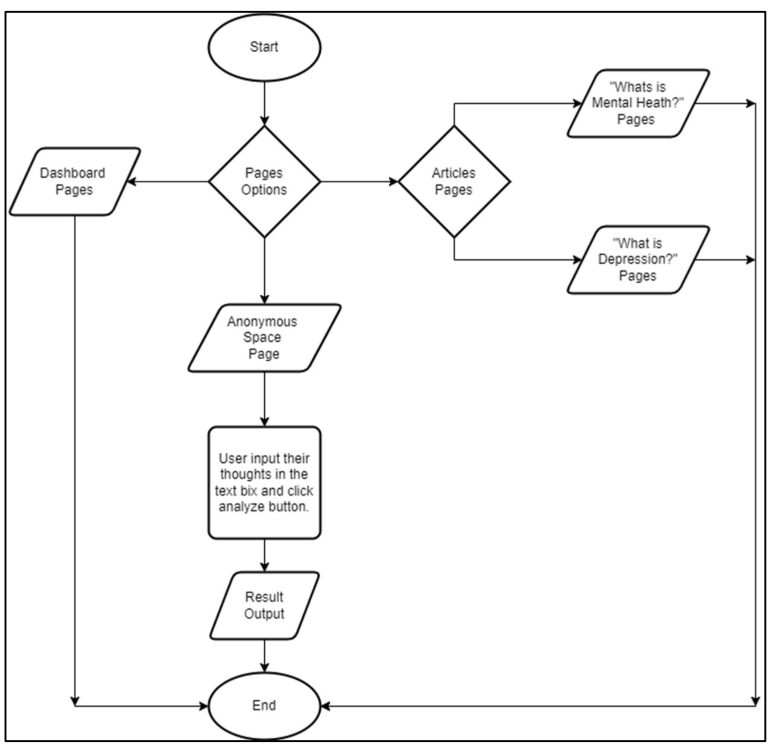


Figure 4. 5 Flowchart of the System

Starting of the flowchart is at the home pages. As shown in the Figure 4.5 above, there will be a “Page Options”, and this will be the “Home Landing Pages” for the websites. Here, user have options to navigate through different pages, including the Home Page, “Anonymous Space Page”, “Articles Page”, and “Dashboard Page”. If the user chooses “Anonymous Space Pages”, the flow in the “Anonymous Space Pages” is as below:

1. Users can freely express their thoughts and feelings through text input.
2. Upon input, they can click the "Analyze" button to have their text evaluated for depression symptoms.
3. The sentiment analysis algorithm processes the text and provides a prediction based on depression indicators.
4. The result is displayed, indicating whether the text exhibits depression symptoms and an emoticon corresponding to the result.

If the user chooses the “Articles Page” in the Home Landing Pages, the flow for the “Articles Pages” is as follow below:

1. This page contains two sub-menus, "What is Mental Health?" and "What is Depression."
2. Users can click on of these sub-menus to access articles that offer information on definitions, tips, symptoms, and relevant videos.
3. If the user clicks “What is Mental Health?”, then the articles about the mental health information will be shows. Same goes with the user clicks “What is Depression?”, then the articles about all the related information about depression will be shown.

If the user chooses the “Dashboard Pages” in the Home Landing Pages, this is the content that will be provided in the ”Dashboard Page”:

1. The Dashboard presents visualizations derived from the sentiment analysis.
2. Word clouds showcase the most common words in the dataset and specific to depression or non-depression labels.
3. A bar chart displays the sentiment distribution.

4. A histogram illustrates the text length distribution for both depression and non-depression labels.

4.3 Back-End Development

This section delves into various aspects of Back End Development involved in the creation of the proposed system. Starting with Data Preprocessing, the steps taken to prepare and organize the data for further analysis are explored. Subsequently, Data Cleaning is addressed, focusing on techniques employed to ensure the data's quality and consistency. Data Vectorization is then discussed, shedding light on the process of transforming textual data into numerical vectors for machine learning analysis. Furthermore, the implementation of the Machine Learning Model takes centre stage as it forms the core component of the Back End, enabling sentiment analysis. Through these stages, Back End Development significantly contributes to the effectiveness and accuracy of the proposed system in addressing mental health sentiment analysis.

4.3.1 Data Pre-processing

Based on the Figure 4.6 below, the 'snscrape' library is used to access Twitter data through the 'sntwitter' module.

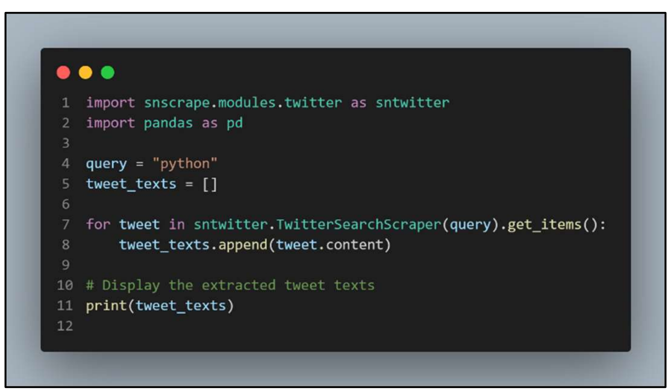


Figure 4. 6 Scaping Data from Twitter Using 'Snscrape' module

The provided code utilizes the 'snscrape' library to access Twitter data using the 'sntwitter' module and retrieve tweets related to the specified query "python." The code initializes an empty list called 'tweet_texts' to store the content of each tweet. It then iterates through the TwitterSearchScraper, where the 'query' variable is used to search for tweets related to the term "python." For each tweet found, the content of the tweet, which represents the text of the tweet, is extracted using 'tweet.content' and appended to the 'tweet_texts' list. The loop continues to search for more tweets until it retrieves relevant data based on the given query. Finally, the code prints the 'tweet_texts' list, which contains the text content of all tweets related to the query "python." This code snippet can be used as a starting point to gather Twitter data for further analysis or sentiment classification related to the topic "python.". The text data is then being save in depress_text.csv file and the sentiment label is labelled as 0.

The next step in data pre-processing is load the data from two CSV files which is “depress_text.csv” and “non_depress.csv” using the pandas library. The code is as shown in the Figure 4.7 below.

```
# Load the dataset
dataframe1 = pd.read_csv("depress_text.csv", index_col=0)
dataframe2 = pd.read_csv("non_depress_text.csv", index_col=0)

# Gabungkan kedua DataFrame
data = pd.concat([dataframe1, dataframe2])

# Reset the index of the DataFrame
data = data.reset_index()
```

Figure 4. 7 Load the dataset

Each CSV file represents a dataset with tweet text related to different sentiments, where "depress_text.csv" contains tweets related to depression (sentiment label 0) and "non_depress_text.csv" contains tweets without depression (sentiment label 1).

The data from both CSV files is read into separate pandas DataFrames, 'dataframe1' and 'dataframe2.' After that, the two DataFrames are merged into a single DataFrame called 'data' using the 'pd.concat()' function, which effectively combines the rows from both DataFrames while keeping the original index of each row.

To ensure proper indexing of the combined data, the 'reset_index()' method is applied to the 'data' DataFrame, which resets the index to consecutive integers starting from 0. This step is crucial to avoid potential indexing issues and ensure a seamless integration of the data from the two datasets.

By merging the datasets and resetting the index, the 'data' DataFrame now contains all the tweet text along with their respective sentiment labels, ready for further data preprocessing and analysis tasks.

4.3.2 Data Cleaning

Figure 4.8 below shows the code for data cleaning:

```python
# Data Cleaning
stop_words = set(stopwords.words('english'))
lemmatizer = WordNetLemmatizer()
stemmer = PorterStemmer()

def clean_text(text):
    # Remove double quotation marks
    text = text.replace('"', '')
    # Remove RT tags
    text = re.sub(r'^RT[\s]+', '', text)
    # Remove user tags (@Usertag)
    text = re.sub(r'@[A-Za-z0-9_]+', '', text)
    # Remove URLs
    text = re.sub(r'https?:\/\/\S+', '', text)
    # Remove hashtags and keep only the text
    text = re.sub(r'#', '', text)
    # Remove emojis
    text = emoji.demojize(text)
    # Remove punctuation
    text = text.translate(str.maketrans('', '', string.punctuation))
    # Convert to lowercase
    text = text.lower()
    # Remove word repetition
    text = re.sub(r'\b(\w+)( \1\b)+', r'\1', text)
    # Handle contractions
    text = contractions.fix(text)
    # Tokenize text
    tokens = nltk.word_tokenize(text)
    # Remove stopwords, perform stemming and lemmatization
    tokens = [lemmatizer.lemmatize(stemmer.stem(word)) for word in tokens if word not in stop_words]
    text = ' '.join(tokens)
    return text

data['Cleaned_Text'] = data['Text'].apply(clean_text)
```

Figure 4. 8 Data Cleaning

The provided code performs data cleaning on a text column named 'Text' in a DataFrame named 'data'. The purpose of data cleaning is to process and standardize the text data, making it more suitable for further analysis, such as sentiment analysis.

Here's an explanation of each step in the data cleaning code:

1. Remove double quotation marks:

   This step removes any double quotation marks from the text as they are not relevant for sentiment analysis.

2. Remove RT tags:

   This step removes "RT" tags, which often appear at the beginning of retweeted messages on Twitter. Removing them helps clean the text from any retweet-related content.

3. Remove user tags (@Usertag):

   User tags, indicated by '@' symbols followed by usernames, are removed from the text. This ensures that user-specific information is not considered during sentiment analysis.
4. Remove URLs:

   Any URLs (web links) present in the text are removed since they do not contribute to sentiment analysis.
5. Remove hashtags and keep only the text: Hashtags (words or phrases preceded by '#') are removed, leaving only the text content for analysis.
6. Remove emojis:

   Emojis are converted into text representations using the 'emoji.demojize()' function. This ensures that emojis do not affect the sentiment analysis process.
7. Remove punctuation:

   All punctuation marks are removed from the text to avoid any interference with the sentiment analysis.
8. Convert to lowercase:

   The entire text is converted to lowercase, ensuring uniformity and consistency during the analysis.
9. Remove word repetition:

   This step removes repeated words (e.g., "happy happy happy") and keeps only one occurrence of each word.
10. Handle contractions:

    Contractions (e.g., "won't" to "will not") are expanded to their full form for consistency and analysis.
11. Tokenize text:

    The text is tokenized into individual words, separating them for further processing.
12. Remove stopwords, perform stemming, and lemmatization: Stopwords (commonly used words with little meaning, such as "the," "and," "is") are removed to reduce noise in the data. Additionally, stemming and lemmatization are applied to normalize words and reduce inflectional forms to their base or root form.

After applying these cleaning steps, the cleaned text is stored in a new column named 'Cleaned_Text' in the DataFrame, ready for further analysis such as sentiment analysis. The cleaned text will have consistent formatting and content, making it easier to extract meaningful insights from the text data.

4.3.3 Data Vectorization

The next step is the data vectorization. Figure 4.9 below shows the code for the implementation of data vectorization and also for the data modelling.

```
# Pipeline for Data Processing and Modeling
pipeline = Pipeline([
    ('tfidf', TfidfVectorizer(ngram_range=(1, 3), min_df=2, max_df=0.8, max_features=5000)),
    ('feature_selection', SelectKBest(score_func=chi2, k='all')),
    ('model', LinearSVC())
])

# Hyperparameter Tuning using GridSearchCV
parameters = {
    'tfidf__ngram_range': [(1, 1), (1, 2), (1, 3)],
    'tfidf__max_features': [1000, 5000, 10000],
    'model__C': [0.1, 1, 10]
}

grid_search = GridSearchCV(pipeline, parameters, cv=5, error_score='raise')
grid_search.fit(X_train, y_train)

# Best Model and Parameters
best_model = grid_search.best_estimator_
best_parameters = grid_search.best_params_
```

Figure 4. 9 Data Vectorization and Modelling

The data vectorization step is performed as part of the pipeline within the code. Data vectorization is a process of converting the raw text data into a numerical representation that machine learning algorithms can process and analyze.

Here's how the data vectorization step is performed in the code. The first step is separate the features (Text) and labels (Sentiment). The original data consists of two main components: the 'Cleaned_Text' column containing the pre-processed text data (features) and the 'Sentiment' column containing the corresponding sentiment labels (labels). The 'X' variable holds the features (cleaned text), and the 'y' variable holds the labels (sentiments).

Figure 4.10 below shows the code for separating the features (text) and label(sentiment).

```python
# Separate the features (text) and labels (sentiment)
X = data['Cleaned_Text']
y = data['Sentiment']
```

Figure 4. 10 Separate the features (text) and label (sentiment)

Then the next step is data splitting. Figure 4.11 below shows the code for the data splitting.

```python
X_train, X_test, y_train, y_test = train_test_split(X,
                                                    y,
                                                    test_size=0.2,
                                                    random_state=42)
```

Figure 4. 11 Data Splitting

The dataset is split into training and testing sets using the 'train_test_split' function from scikit-learn. The training set (X_train, y_train) is used to train the machine learning model, and the testing set (X_test, y_test) is used to evaluate its performance.

After that, proceed to the next step which is pipeline for Data Processing and Modeling as shown in the Figure 4.9. A scikit-learn pipeline is created to streamline the data processing and modeling steps. Based on the Figure 4.6, the pipeline consists of three main components, but in this step only the first two of the main components will be explained since its related to the data vectorization which is text vectorization and feature selection:

1. Text Vectorization:

- TfidfVectorizer: This step transforms the raw text data into a numerical representation using the TF-IDF (Term Frequency-Inverse Document Frequency) technique. It converts the text into a matrix where each row represents a document, and each column represents a unique word or n-gram (a sequence of consecutive words).
- Parameters: The TfidfVectorizer is configured with parameters such as the n-gram range (1 to 3), which defines the length of the word sequences to consider, the minimum document frequency (2), and the maximum document frequency (0.8) to filter out very common or rare words. The maximum number of features (5000) limits the total number of unique words in the vocabulary.

2. Feature Selection:

- SelectKBest: This step selects the top K best features based on statistical measures, in this case, the chi-squared (chi2) test. It helps to reduce the dimensionality of the data and focus on the most informative features.
- Parameters: The SelectKBest is set to select 'all' features, meaning it keeps all the features selected by the chi-squared test.

4.3.4 Machine Learning Model

After data vectorization has been done, the next step in the process is to apply a data modelling algorithm which uses SVM with LinearSVC, for sentiment analysis.

```python
# Pipeline for Data Processing and Modeling
pipeline = Pipeline([
    ('tfidf', TfidfVectorizer(ngram_range=(1, 3), min_df=2, max_df=0.8, max_features=5000)),
    ('feature_selection', SelectKBest(score_func=chi2, k='all')),
    ('model', LinearSVC())
])

# Hyperparameter Tuning using GridSearchCV
parameters = {
    'tfidf__ngram_range': [(1, 1), (1, 2), (1, 3)],
    'tfidf__max_features': [1000, 5000, 10000],
    'model__C': [0.1, 1, 10]
}

grid_search = GridSearchCV(pipeline, parameters, cv=5, error_score='raise')
grid_search.fit(X_train, y_train)

# Best Model and Parameters
best_model = grid_search.best_estimator_
best_parameters = grid_search.best_params_

print("Best Model:", best_model)
print("Best Parameters:", best_parameters)

# Make Predictions on the Test Set
y_pred = best_model.predict(X_test)
```

Figure 4. 12 Machine Learning Model

From the Figure 4.12 above, this code represents the implementation of a pipeline for data processing and modeling in a sentiment analysis task. Below is an explanation of the process of Model Training.

1. Based on the Figure 4.12, one of the three main components of the pipeline is related with the data modelling which is LinearSVC model. The LinearSVC is a variant of the Support Vector Machine (SVM) algorithm that performs linear classification. It is used as the machine learning model for sentiment analysis. The SVM algorithm aims to find an optimal hyperplane that best separates data points belonging to different classes (sentiments) in the vectorized feature space.
2. Hyperparameter Tuning using GridSearchCV:

To identify the best hyperparameters for the TfidfVectorizer and LinearSVC, hyperparameter tuning is performed using GridSearchCV. GridSearchCV performs an exhaustive search over a specified parameter grid and selects the combination of hyperparameters that results in the best performance based on cross-validated evaluation.

The parameters dictionary contains different hyperparameter options for 'tfidf__ngram_range' (word combinations), 'tfidf__max_features' (maximum features), and 'model__C' (SVM regularization parameter).

3. Best Model and Parameters:

   After performing hyperparameter tuning, the best model and corresponding best parameters are obtained using the best_estimator_ and best_params_ attributes of the GridSearchCV object, respectively.

4. Make Predictions on the Test Set:

   Finally, the best model obtained from hyperparameter tuning is used to make predictions on the test set (X_test). The predictions (y_pred) are then compared to the true labels (y_test) to evaluate the model's performance in classifying sentiments correctly.

By combining the data processing steps with the LinearSVC model and tuning the hyperparameters using GridSearchCV, the code effectively implements a sentiment analysis model for predicting sentiments (positive or negative) based on the text data.

4.4 Front-End Development

Front End Development is a crucial aspect of building a web application as it focuses on creating the user interface that users interact with directly. This section explores two fundamental components of Front-End Development for our proposed system. First, is the User Interface Design, where the process of designing an appealing and user-friendly interface is discussed, ensuring seamless interaction with the system. Next is the explanation for each pages in the mental health websites, showcasing the functionalities and tools available to users for effective engagement with the mental health sentiment analysis platform. Emphasizing these Front-End aspects aims to provide users with a welcoming and supportive platform for mental health awareness and analysis.

4.4.1 User Interface Design

This section explores the design of several key pages that constitute the user interface for the proposed system. The Home Page serves as an inviting entry point, providing an overview of the platform's purpose and features. Next, the Anonymous Space Page offers users a confidential environment to express their thoughts and feelings related to mental health. The Articles Page presents valuable information, featuring dedicated sections explaining "What is Mental Health?" and "What is Depression?". Lastly, the Dashboard Page serves as a personalized hub for users, providing insightful visualizations and analysis of their mental health sentiment data.

4.4.2 Home Pages

The Home Page serves as the initial landing page when users access the web application. It is the first page that users encounter, and its primary function is to welcome and guide them through the platform. Figure 4.13 and Figure 4.14 below shows the home pages for this website.

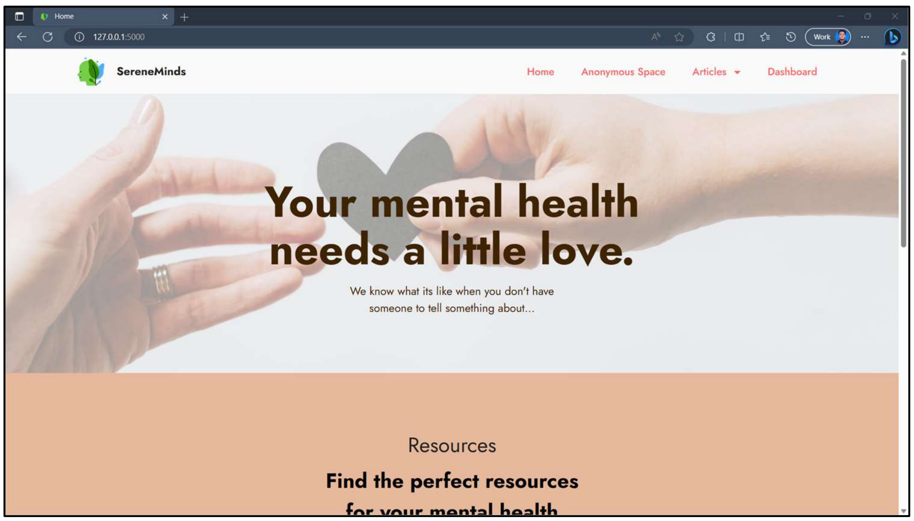


Figure 4. 13 Home Page 1

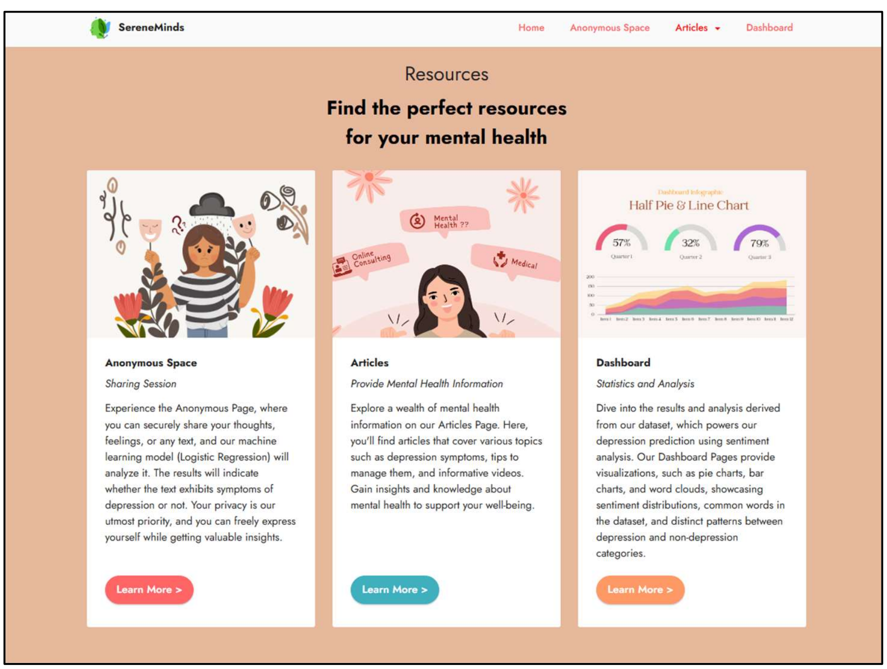


Figure 4. 14 Home Page 2

From the Figure 4.14 above, there are three essential features found in the home pages which is “Anonymous Space”, “Articles” and “Dashboard”.

4.4.3 Anonymous Space Pages

Figure 4.15, Figure 4.16 and 4.17 shows the user interface design for the “Anonymous Space” page.

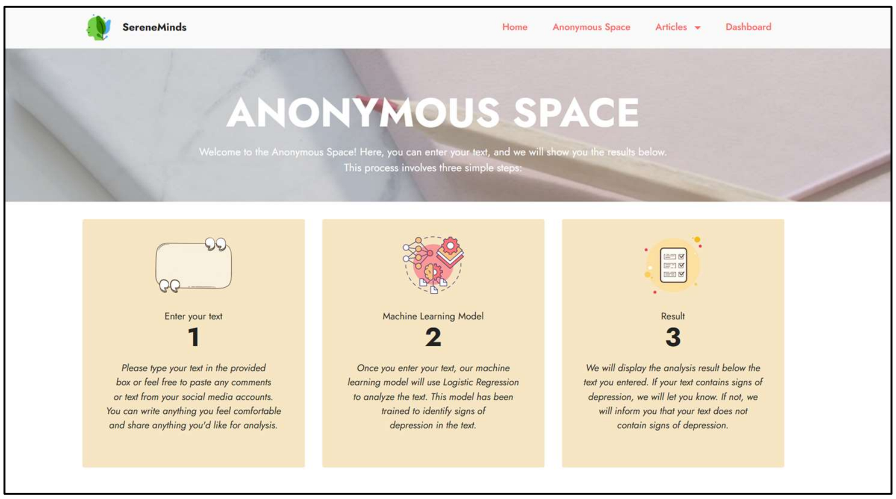


Figure 4. 15 Anonymous Space Page 1

The Anonymous Space Page is a dedicated section within the web application that allows users to freely express their feelings, thoughts, and emotions through text without disclosing their identity. This space provides a safe and confidential environment for users to share their experiences related to mental health, seek support, or simply express themselves without the fear of judgment.

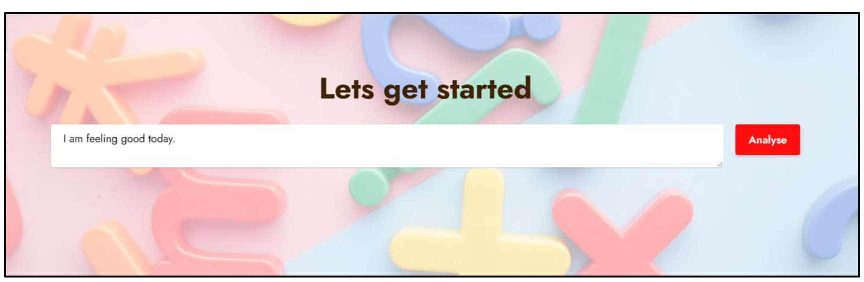


Figure 4. 16 Anonymous Page 2

When users submit their text in the Anonymous Space as shown in the Figure 4.16 above, the platform performs a sentiment analysis on the text. Sentiment analysis is a process that examines the text's emotional tone and determines whether it exhibits symptoms of depression or not. The analysis considers various linguistic patterns and sentiment indicators to assess the text's emotional state.

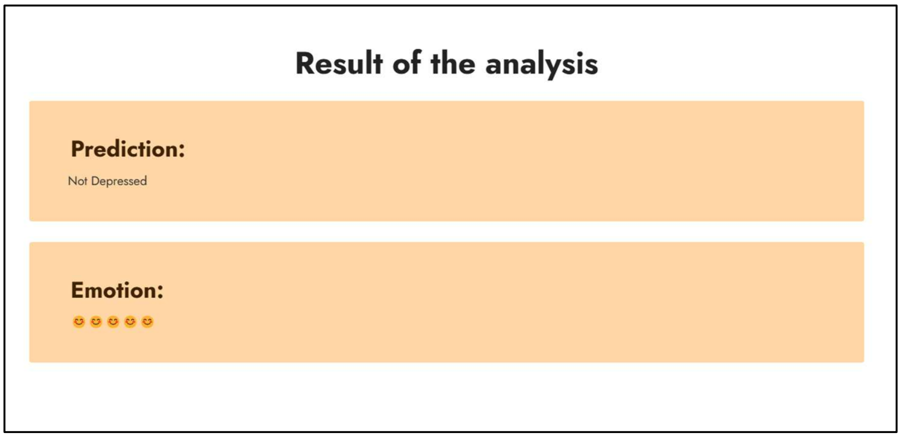


Figure 4. 17 Anonymous Space Page

The result of the sentiment analysis is then displayed below as shown in the Figure 4.17, below the user's submitted text. The platform provides feedback to the user, informing them whether the text shows any signs of depression or not. If the analysis identifies concerning patterns or indicates potential symptoms of depression, the platform may encourage users to seek professional help or provide resources for mental health support.

4.4.4 Articles Pages

The "Articles" section of the web application consists of two informative submenus: "What is Mental Health?" and "What is Depression?". These submenus serve as valuable resources for users seeking information about these essential mental health topics.

The user interface design for the first articles of "What is Mental Health?" article, is as shown in the Figure 4.18, Figure 4.19, Figure 4.20, Figure 4.21, and Figure 4.22.

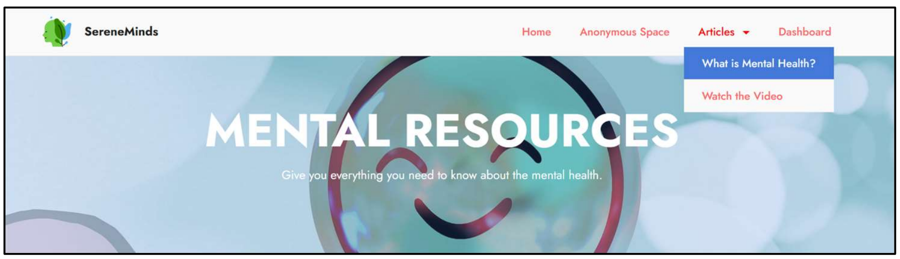


Figure 4. 18 Articles Page “What is Mental Health?” 1

Based on Figure 4.19, this submenu provides a comprehensive overview of mental health, offering users a clear definition of what mental health encompasses. It discusses the importance of mental well-being and emphasizes the significance of taking care of one's mental health.

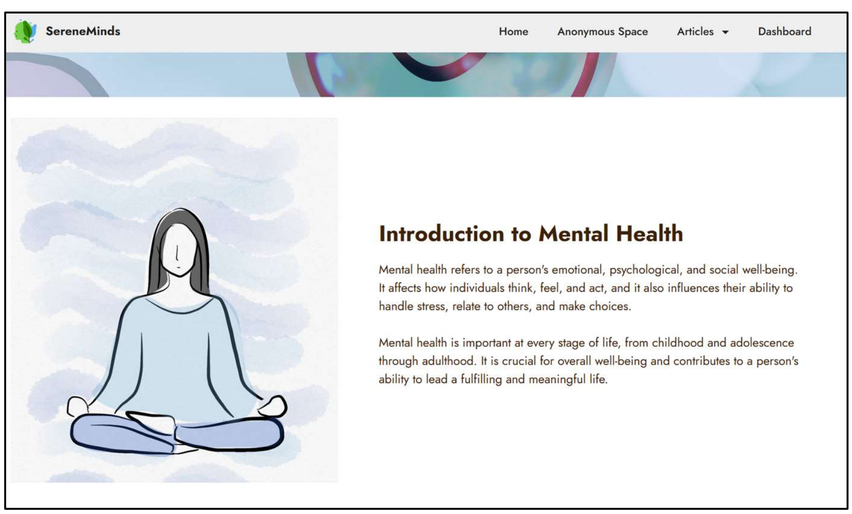


Figure 4. 19 Articles Page “What is Mental Health?” 2

The section also includes practical tips and strategies for maintaining good mental health. These tips may include practices for stress management and ways to foster emotional resilience.

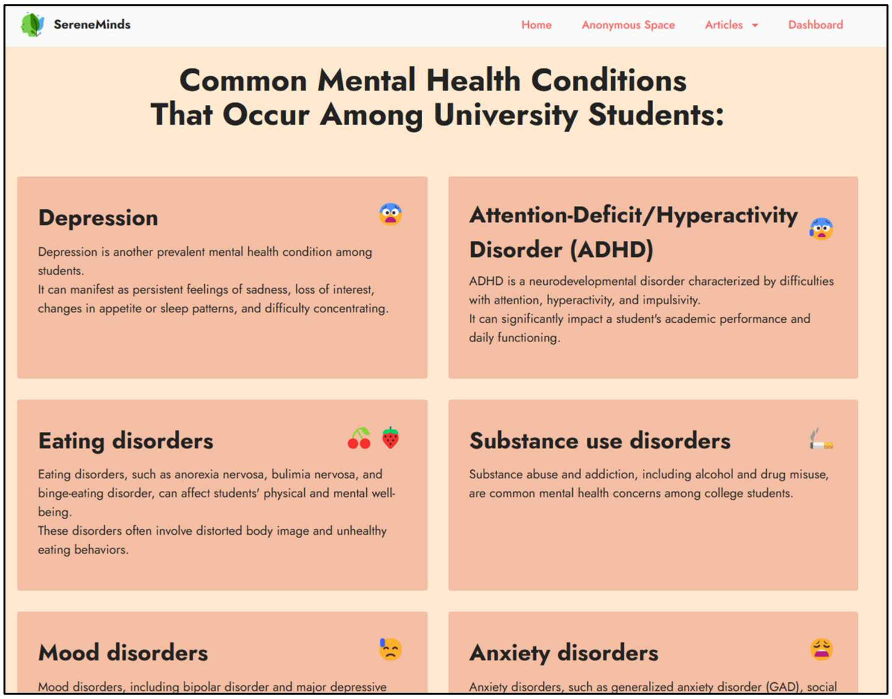


Figure 4. 20 Articles Page “What is Mental Health?” 3

Additionally, this submenu highlights common mental health conditions that occur among university students, such as depression, eating disorder, and mood disorders, as shown in the Figure 4.20 above. Figure 4.21 below show the symptoms and warning signs for the mental health conditions.

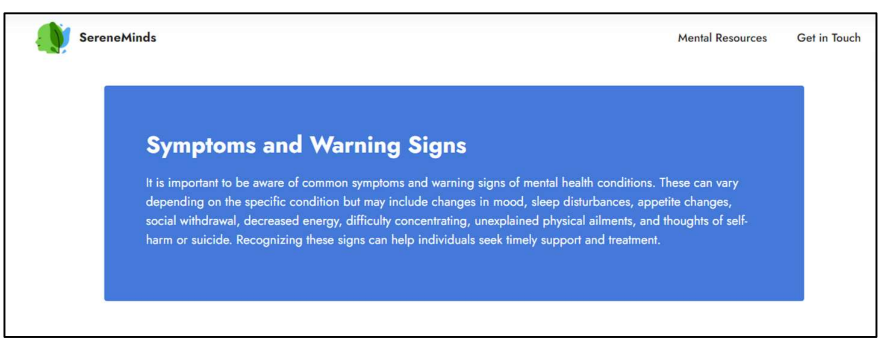


Figure 4. 21 Articles Page “What is Mental Health?” 4

To enhance users' understanding, the "Articles" section may provide informative videos related to mental health, real-life experiences that shed light on the importance of mental well-being like shown in the Figure 4.20 below.

Figure 4. 22 Articles Page “What is Mental Health?” 5

Next is the user interface design for the second articles page which is "What is Depression?", as shown in the Figure 4.23, Figure 4.24, Figure 4.25, Figure 4.26, Figure 4.27, Figure 4.28 and Figure 4.29.

From the Figure 4.23, Figure 4.24, Figure 4.25, it shows the delves into the definition and characteristics of depression, educating users about this prevalent mental health condition.

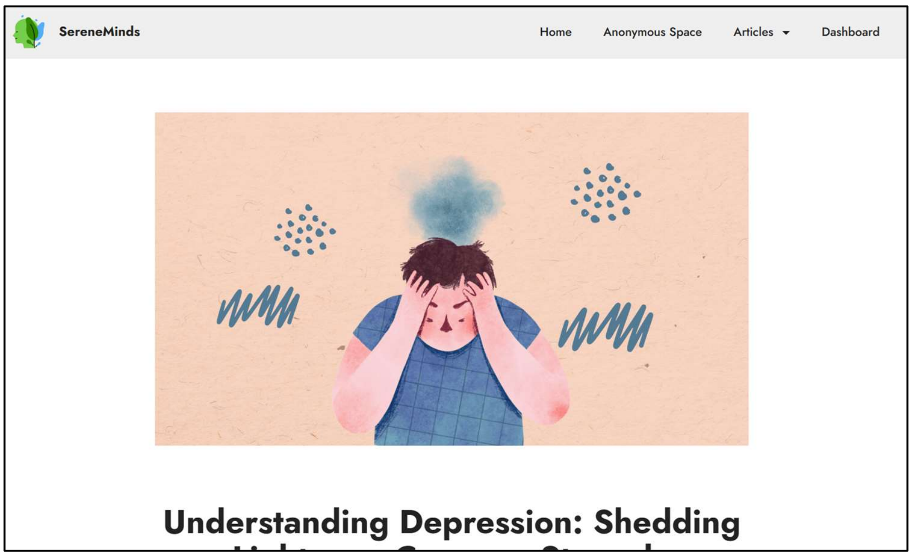


Figure 4. 23 Articles Page "What is Depression?" 1

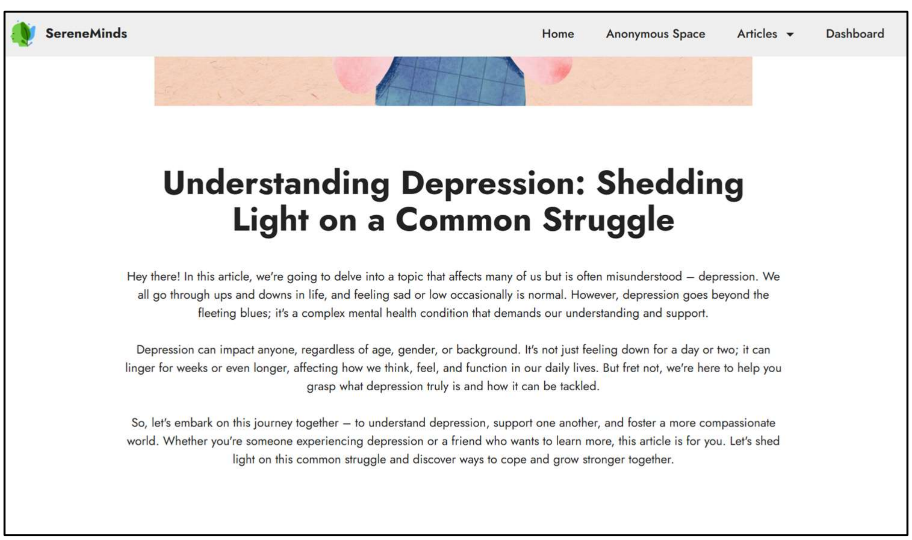


Figure 4. 24 Articles Page 7 "What is Depression?" 2

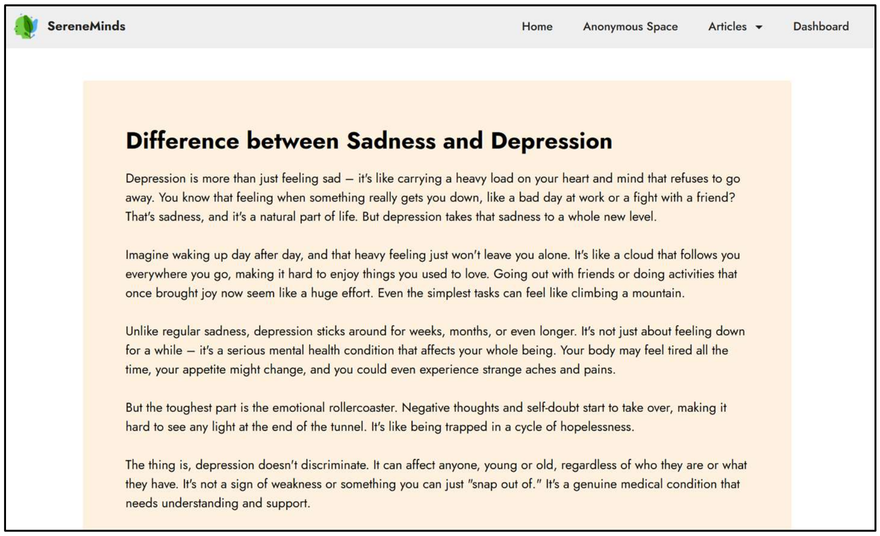


Figure 4. 25 Articles Page "What is Depression?" 3

In the "What is Depression?" page, there are three buttons available for users to access specific information within the article as shown in the Figure 4.26. Each button corresponds to a particular section or block of content related to depression.

Figure 4. 26 Articles Page "What is Depression?" 4

- Symptoms Button:
  When the user clicks on the "Symptoms" button, it will automatically redirect them to the section of the article as shown in the Figure 4.27, that provides detailed information about the symptoms of depression. This section may include a comprehensive list of common symptoms experienced by individuals who are dealing with depression. It may also describe how these symptoms can manifest in various aspects of life, such as emotions, behaviour, and physical well-being.

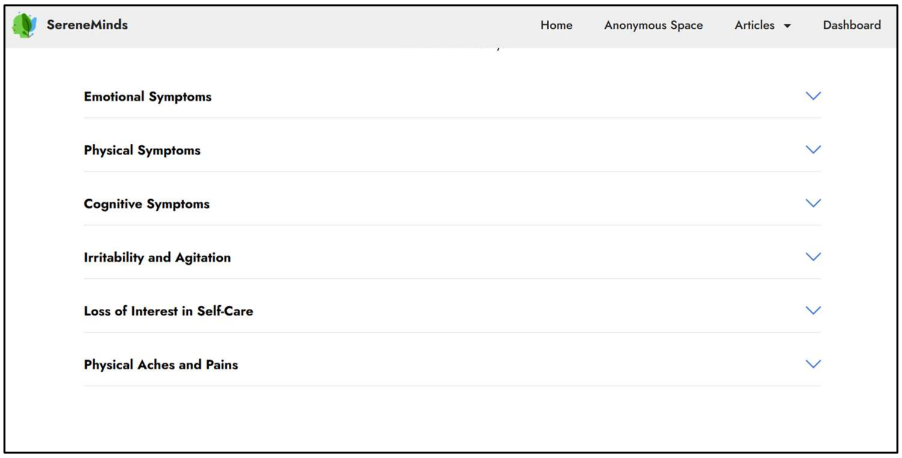


Figure 4. 27 Articles Page "What is Depression?" 5

- Depression Tips Button:

  Clicking on the "Depression Tips" button will take users directly to the section as shown in the Figure 4.28, that offers tips and coping strategies for dealing with depression. This section may provide valuable advice on managing depressive feelings, seeking support from loved ones or professionals, and engaging in self-care practices to promote mental well-being.

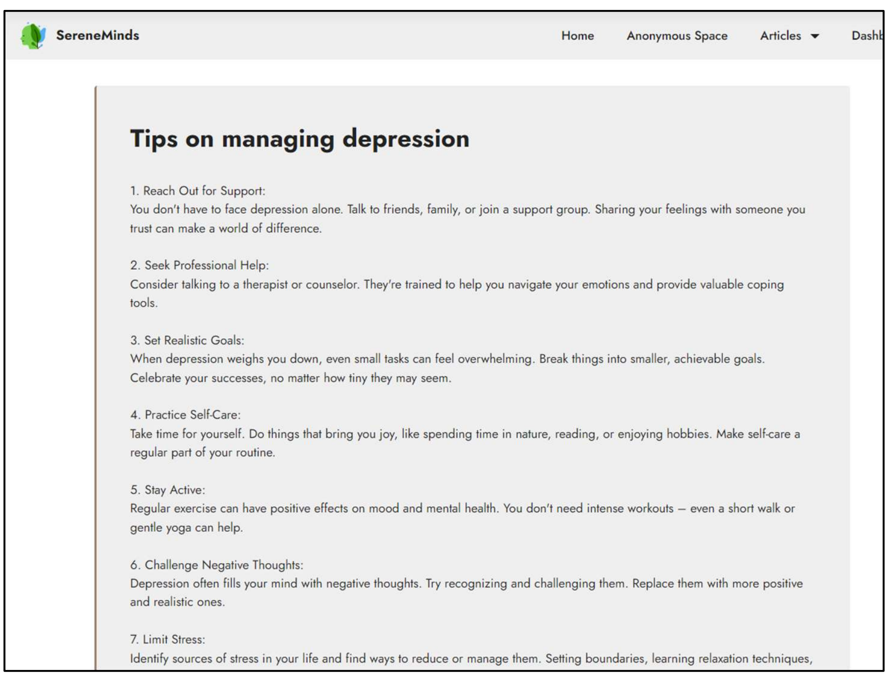


Figure 4. 28 Articles Page "What is Depression?" 6

- Watch Videos Button:

  Upon clicking the "Watch Videos" button, users will be redirected to a section as shown in the Figure 4.29, that features relevant and informative videos related to depression. These videos may include personal stories of individuals sharing their experiences with depression, expert insights on understanding and managing depression, or therapeutic resources like guided meditation or relaxation techniques.

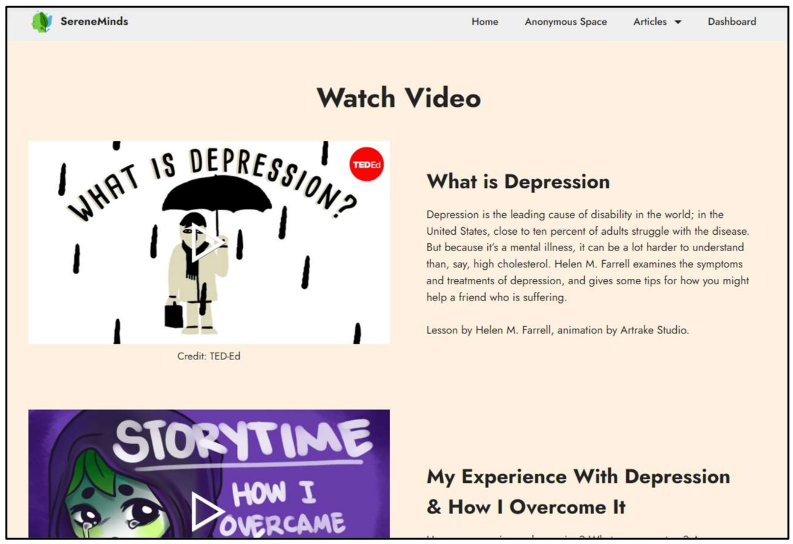


Figure 4. 29 Articles Page "What is Depression?" 7

By providing in-depth information, tips, and engaging videos, the "Articles" section aims to empower users with knowledge about mental health and depression. It serves as an educational resource, fostering mental health literacy and promoting positive mental well-being among users.

4.4.5 Dashboard Pages

The Dashboard Pages are designed to visually showcase the valuable insights derived from the sentiment analysis model, giving users a sneak peek into the powerful data visualization and analysis capabilities available to them. Through attractive visualizations like charts and word clouds, users can easily understand the processed mental health sentiment data. Figure 4.30 below shows the overview of the dashboard pages in the websites.

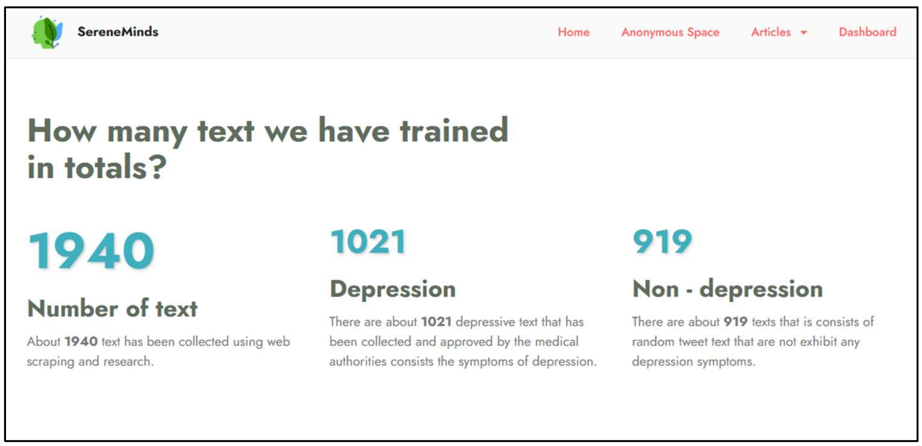


Figure 4. 30 Dashboard Page 1

Figure 4.31 and Figure 4.32 below is the Wordclouds that display the most frequently used words in the dataset in a captivating way, highlighting common themes and topics.

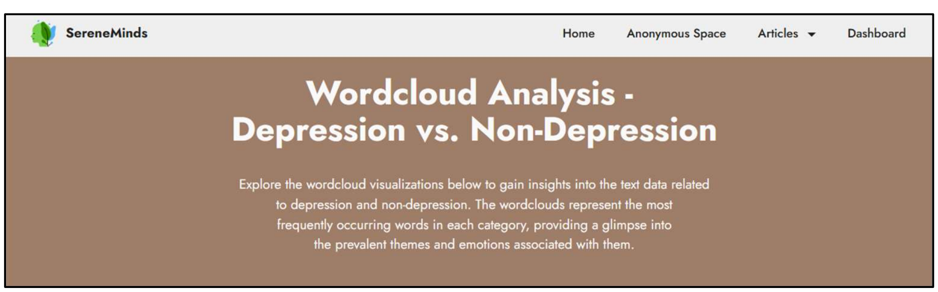


Figure 4. 31 Dashboard Page 2

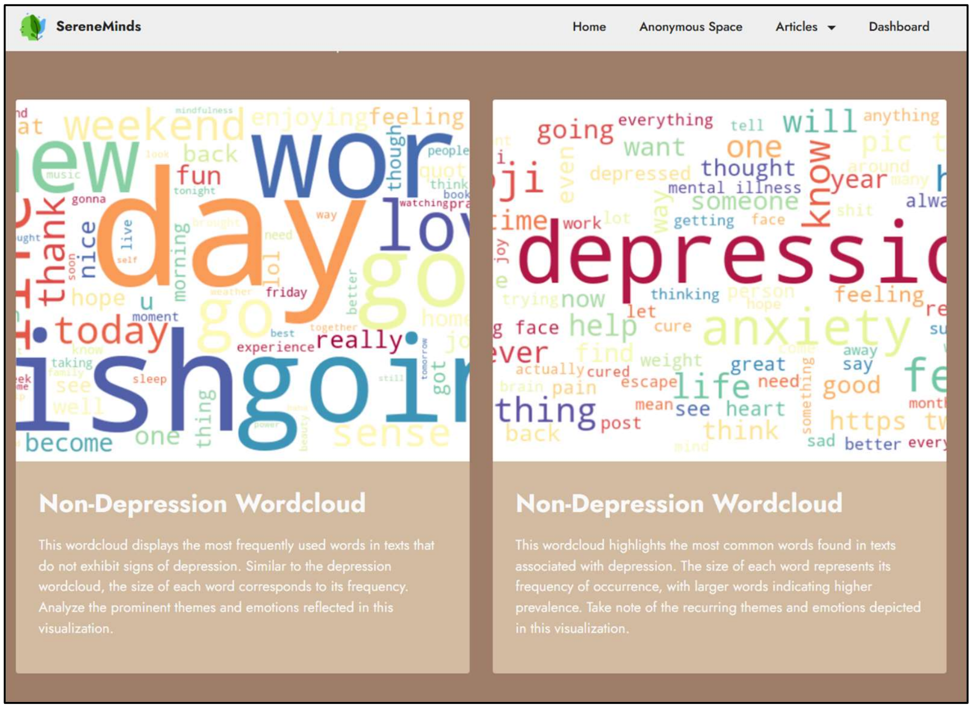


Figure 4. 32 Dashboard Page 3

Figure 4.33 shows the pie chart that visually illustrates the distribution of sentiments, revealing the emotional tone of the analyzed text. These visualizations encourage users to explore more insights in the full Dashboard Pages, accessible through a link or button.

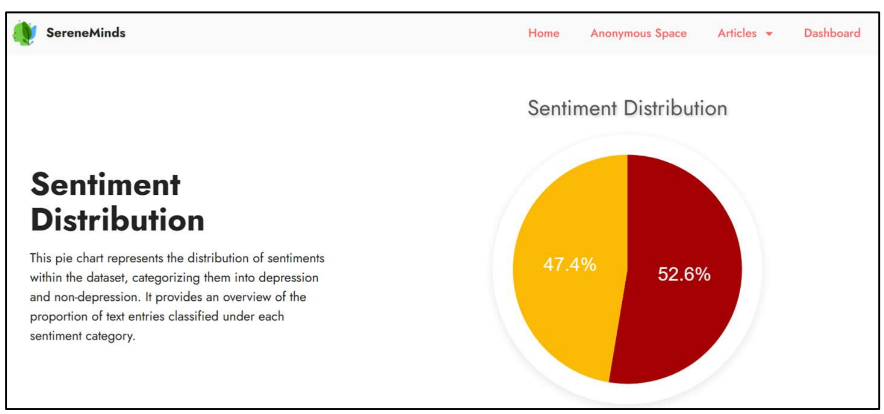


Figure 4. 33 Dashboard Page 4

The Dashboard Pages also provide users with a comprehensive view of the mental health sentiment analysis results, including a unique Text Length

Histogram as shown in the Figure 4.34, created with Plotly.js. This engaging histogram visualizes the distribution of text lengths in the dataset for both depression and non-depression labels.

By employing Plotly.js, users can interactively explore the histogram, gaining valuable insights into the distribution of text lengths. The histogram showcases the frequency of different text lengths, providing a clear understanding of how the length of text data varies within each sentiment category.

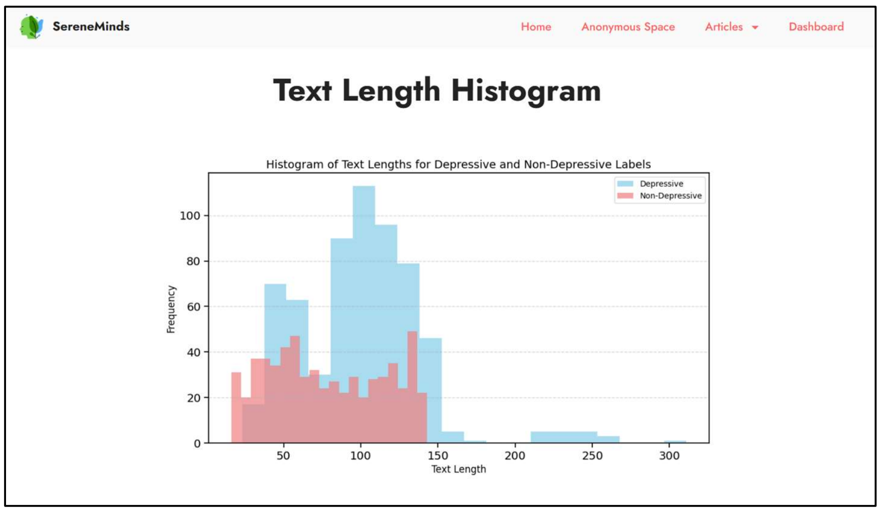


Figure 4. 34 Dashboard Page 5

The bar chart in the Figure 4.35, Figure 4.36 and Figure 4.37 presents the most frequent words in the dataset, in depression column and non-depression column, offering a clear view of their occurrence. Comparing common words within depression and non-depression labels helps users notice any language patterns associated with each sentiment category.

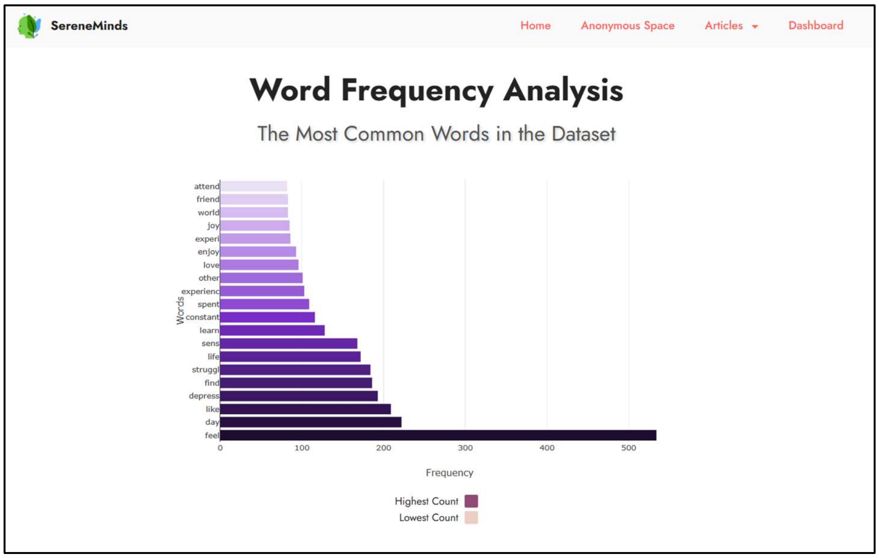


Figure 4. 35 Dashboard Page 6

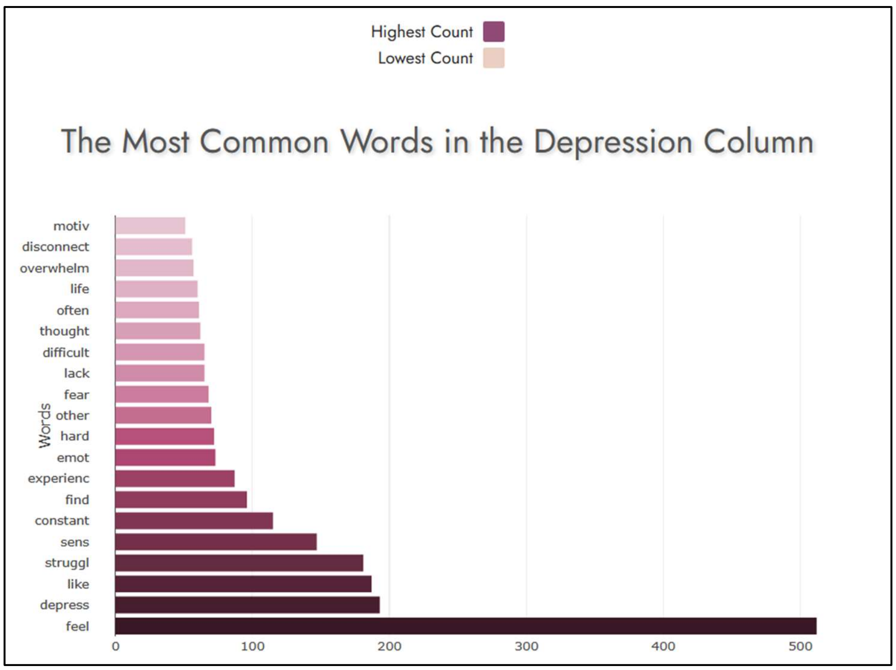


Figure 4. 36 Dashboard Page 7

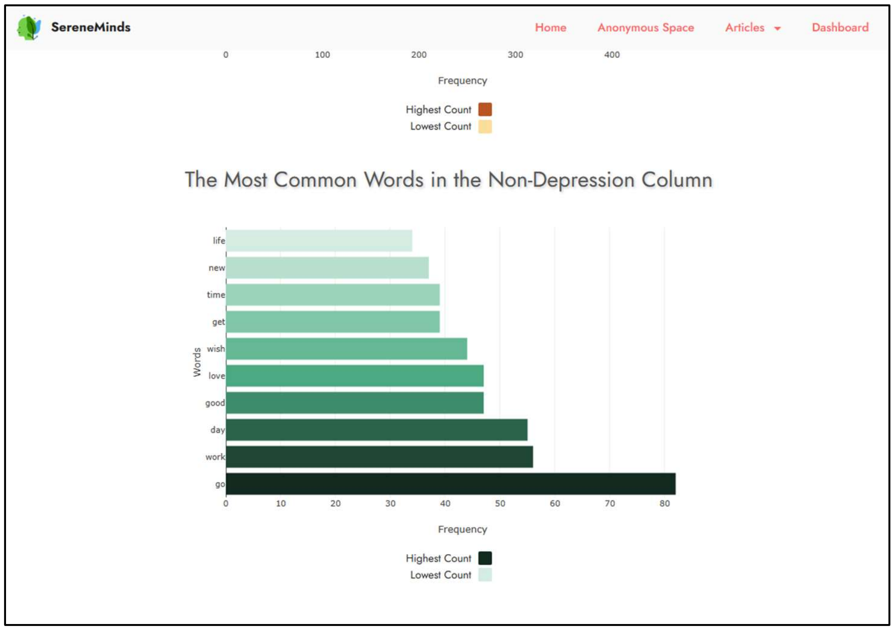


Figure 4. 37 Dashboard Page 8

The Dashboard Pages empower users to monitor their mental health sentiment, identify patterns, and enhance self-awareness. With user-friendly and captivating visualizations, this platform becomes a valuable resource for mental health awareness and support, encouraging users to actively engage in their well-being.

CHAPTER 5

RESULT AND ANALYSIS

The findings and results of the study done for this project are presented in this chapter, which also includes text, tables, figures, and charts presenting the data that was gathered. The research findings are thoroughly analysed in this chapter, together with descriptive explanations, interpretations, and insights.

5.1 Sentiment Analysis Results

This section explains how the sentiment analysis is conducted using the machine learning model that has been created. The analysis is performed on text data obtained from users who share their thoughts and feelings anonymously on the web application. When a user clicks the "Analyze" button after sharing their text, the sentiment analysis model processes the input and provides an output prediction.

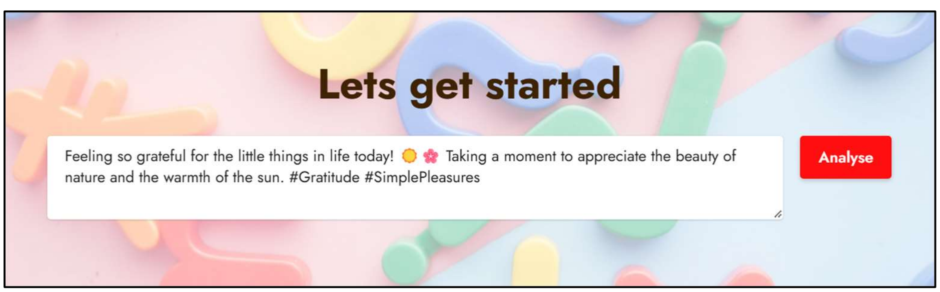


Figure 5. 1 Non-Depression Text Example

Based on the Figure 5.1 above, in the given example text, the user expresses feelings of gratitude and appreciation for the simple pleasures in life, such as the beauty of nature and the warmth of the sun. The text also includes positive emojis, such as “🌞” (sun) and “🌸” (flower), indicating a cheerful and contented mood.

The sentiment analysis process starts with data preprocessing, where the text is cleaned to remove URLs, emojis, punctuation, and other irrelevant information. The text is then converted to lowercase and tokenized into individual words.

Next, the text is vectorized using the Term Frequency-Inverse Document Frequency (TF-IDF) technique, which transforms the words into numerical features suitable for machine learning. The TF-IDF vectorizer assigns higher weights to rare words like "grateful," "little things," and "simple pleasures," which are significant indicators of positive sentiment in the context of this text.

The vectorized text is then fed into the Support Vector Machine (SVM) model with LinearSVC for sentiment classification. The model learns from the labelled training data and predicts whether the input text exhibits depression symptoms or not.

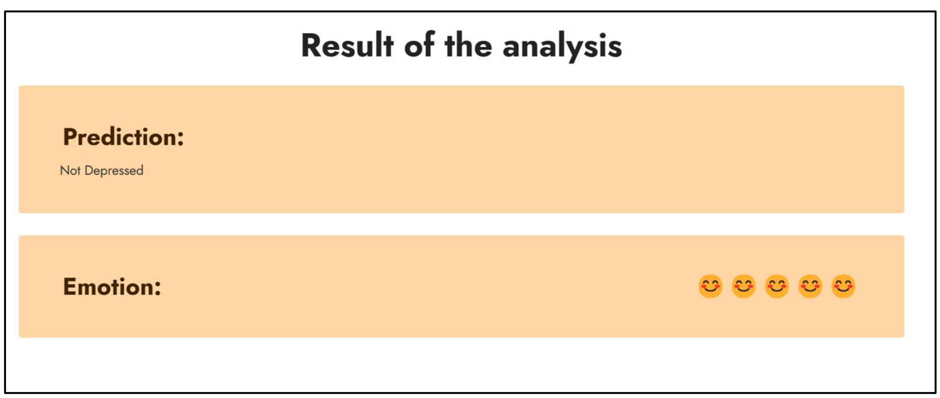


Figure 5. 2 Result of the Analyzation 1

Based on the Figure 5.2 above, the sentiment analysis model predicts that the user's text does not exhibit depression symptoms. The output result is "Not Depressed," which indicates a positive sentiment in the text. The emotion icon “😊 😊 😊 😊 😊” is shown alongside the result, representing the cheerful and positive mood conveyed by the text.

In the Figure 5.3 below, it shows in the given example the user expresses feelings of struggle and sadness, mentioning that the weight of sadness feels unshakeable, and they find it hard to experience joy in anything anymore. The tone of the text is characterized by a sense of hopelessness and emotional burden.

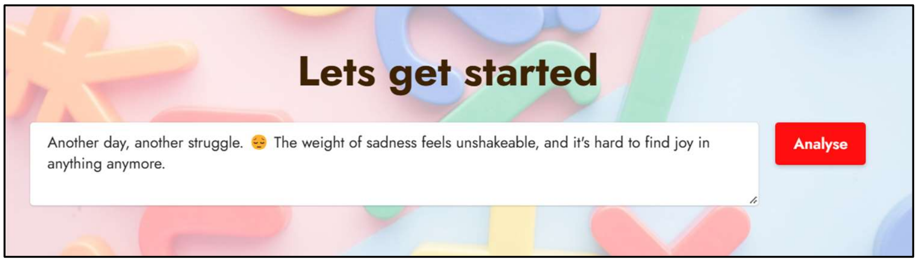


Figure 5. 3 Depression Text Example

As shown in the Figure 5.4 below, the sentiment analysis process predicts that the user's text exhibits symptoms of depression. The output result is "Depression," indicating the presence of depressive sentiments in the text. The emotion icon "😣 😣 😣 😣 😣" is displayed alongside the result, representing the feelings of sadness and distress conveyed by the text.

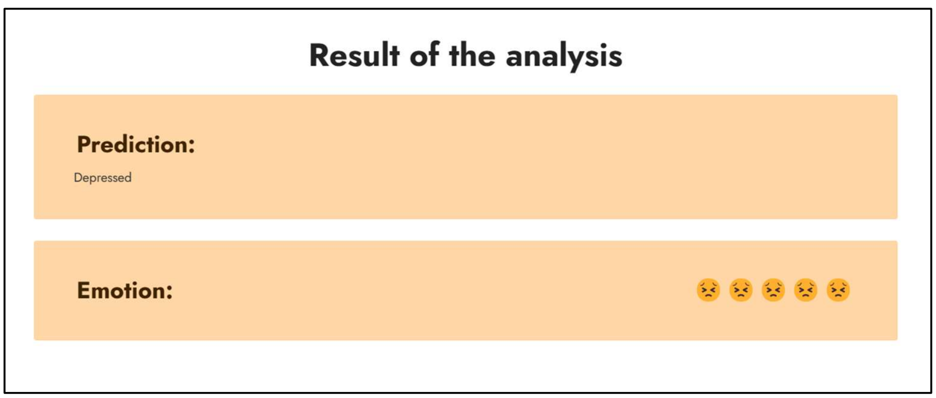


Figure 5. 4 Result of the Analyzation 2

The web application's sentiment analysis accurately recognizes the user's emotional state, highlighting the importance of mental health awareness and support for individuals experiencing such feelings. By providing insights into their emotional well-being, the application encourages users to seek appropriate help and resources to address their emotional struggles.

5.2 Data Visualization

The Data Visualization section in this report employs various engaging and informative techniques to present the results of the sentiment analysis and processed mental health data. Through visually appealing elements, it delves into different aspects of the dataset, offering insights into the text contents and sentiment distribution.

5.2.1 WordCloud

The introduction to the Wordcloud subsection highlights the visual representation of frequently occurring words in the dataset, specifically focusing on depression and non-depression labels. Wordclouds present an intuitive and engaging display of common themes and prominent words within the text data.

This section includes two distinct wordclouds: one for the depression label and the other for the non-depression label.

1. Depression WordCloud
   From the Figure 5.5 below, the word cloud shows that the words 'feel' and 'depress' are the most prominent and frequently occurring words in the depressive sentiment text data. This suggests that these words are highly associated with the expressions and experiences of individuals expressing depressive sentiments.

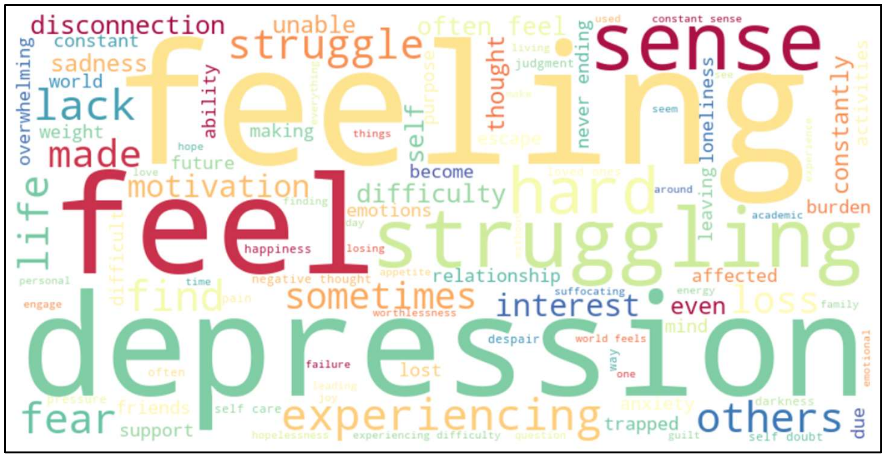


Figure 5. 5 Depression WordCloud

The size of each word in the word cloud indicates its frequency in the text data, with larger words representing higher frequencies. Therefore, the larger appearance of 'feel' and 'depress' indicates that these words are used more frequently in the text data labelled as depressive sentiment.

By analyzing the word cloud, we can gain insights into the common themes and language used in texts expressing depressive sentiments. This information can be valuable in understanding the emotional state and experiences of individuals who exhibit depressive sentiments.

2. Non-depression WordCloud

The word cloud for the non-depress sentiment represents the most frequently occurring words in the text data labelled as non-depress sentiment.

Figure 5. 6 Non-depression WordCloud

From the Figure 5.6 above, the words 'spent' and 'life' are identified as the most prominent and frequently occurring words in the non-depress sentiment text data.

The appearance and size of each word in the word cloud indicate its frequency in the text data, with larger words indicating higher frequencies. Thus, the larger size of 'spent' and 'life' suggests that these words occur more frequently in the text data labelled as non-depress sentiment.

Analyzing the word cloud can provide insights into the common topics and language used in texts expressing non-depress sentiment. In this case, the presence of 'spent' and 'life' indicates that discussions related to time allocation and overall life experiences are prevalent in the non-depress sentiment text data.

5.2.2 Sentiment Distribution

The sentiment distribution of pie chart in the Figure 5.7 represents the distribution of sentiments within the dataset, categorizing the text entries into two categories: depression and non-depression. The chart provides below is an overview of the proportion of text entries classified under each sentiment category.

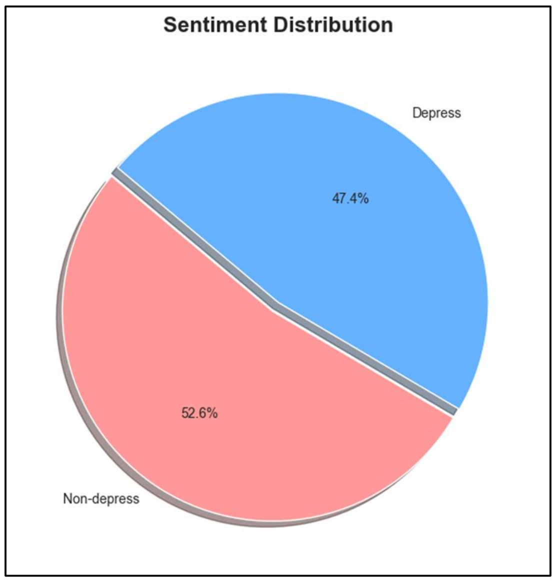


Figure 5. 7 Sentiment Distribution

According to the pie chart from Figure 5.7, the dataset is divided into two main categories: non-depression and depression. The non-depression category accounts for 919 entries, which represents approximately 47.4% of the dataset. On the other hand, the depression category consists of 1,021 entries, also representing approximately 47.4% of the dataset.

The pie chart visually illustrates the balance between the non-depression and depression sentiments within the dataset. It shows that both categories have a similar proportion, indicating that the dataset contains a relatively equal distribution of text entries expressing non-depressive and depressive sentiments. This information can be valuable for further analysis, modelling, and understanding the overall sentiment patterns and trends present in the dataset.

5.2.3 Most Common Word

The bar chart in Figure 5.8 below represents the most common words in the dataset, showing the frequency of occurrence for each word. The chart provides valuable insights into the prevalent words and their respective frequencies, giving an understanding of the key terms that appear frequently in the text data.

.

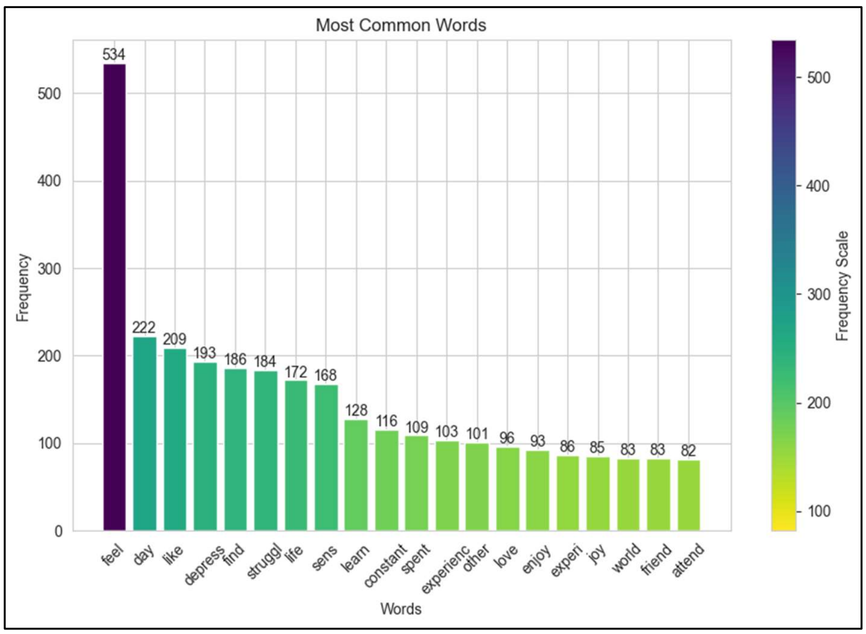


Figure 5. 8 Most Common Words

The chart displays the words on the x-axis and their corresponding frequencies on the y-axis. Each bar represents a word, and its height indicates the frequency of occurrence in the dataset. The taller the bar, the higher the frequency of the word in the text data.

Based on the chart on Figure 5.8, the most common word is "feel," with a frequency of 534. This suggests that the word "feel" appears frequently in the dataset. Other highly frequent words include "day" (222), "like" (209), "depress" (193), and "find" (186), indicating their significance and prevalence in the text data.

The bar chart allows us to identify the most common words in the dataset and gain insights into the topics or themes represented in the text. It helps in understanding the language patterns and prominent keywords that occur frequently.

5.2.4 Most Common Words for Depression Text

The bar chart in Figure 5.9 below represents the most common words in the dataset specifically for the text labelled as depressive sentiment. It displays the words on the x-axis and their corresponding frequencies on the y-axis.

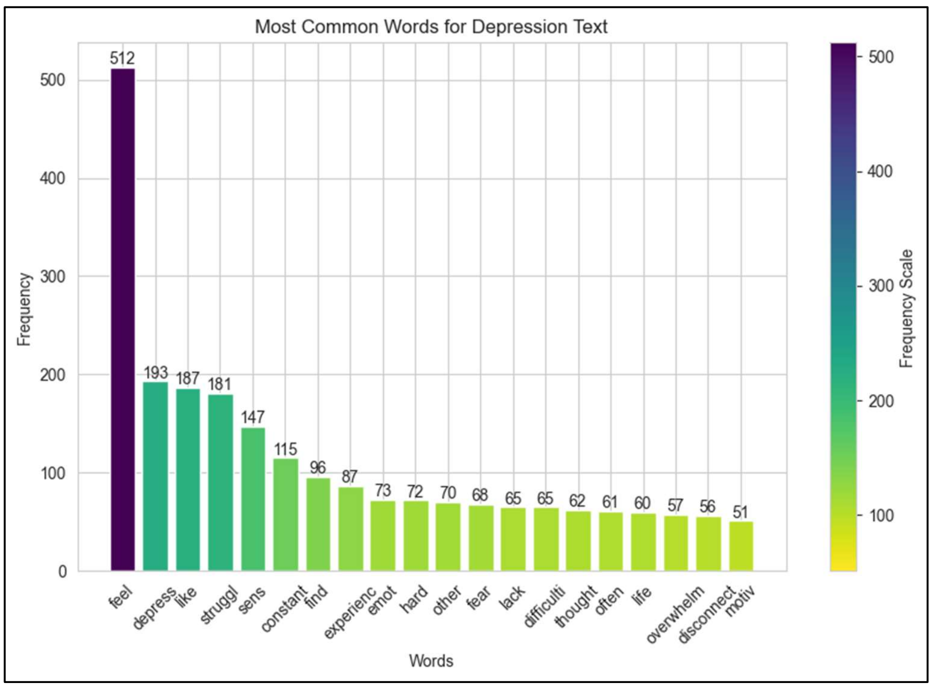


Figure 5. 9 Most Common Words for Depression Text

Based on the chart in Figure 5.9 above, the most common word in the depressive sentiment text is "feel" with a frequency of 512. This suggests that the word "feel" appears frequently in the text labelled as depressive sentiment. Other highly frequent words in the depressive sentiment text include "depress" (193), "like" (187), "struggl" (181), and "sens" (147).

The chart provides insights into the language patterns and prominent keywords that are prevalent in the text labelled as depressive sentiment. These words highlight the emotions, experiences, and challenges associated with depression. By analyzing these common words, one can gain a better understanding of the themes and topics discussed in the text labelled as depressive sentiment.

5.2.5 Most Common Words for Non-Depression Text

The bar chart in Figure 5.10 below represents the most common words in the dataset specifically for the text labeled as non-depressive sentiment. It displays the words on the x-axis and their corresponding frequencies on the y-axis.

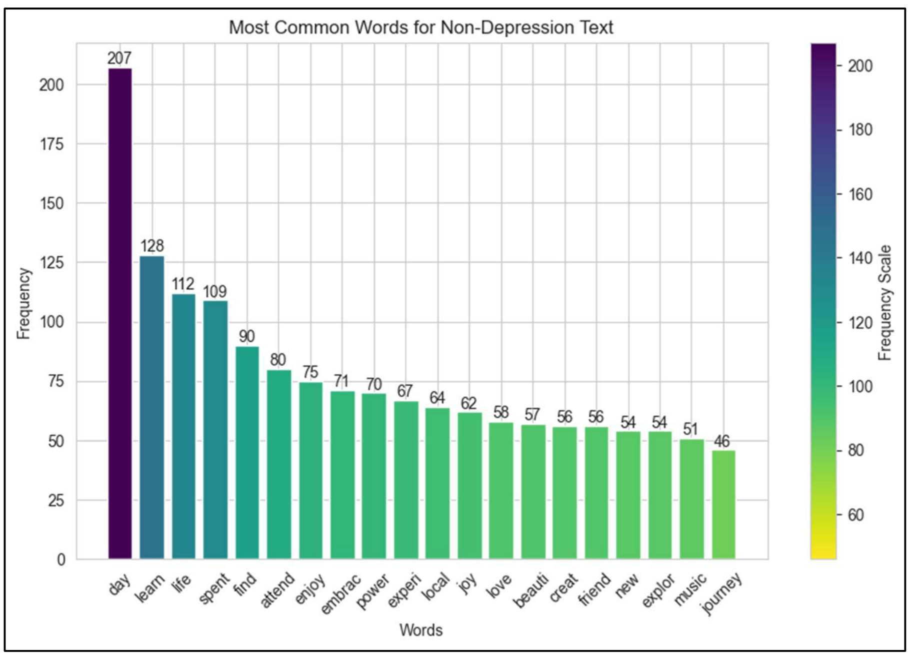


Figure 5. 10 Most Common Words for Non-Depression Text

Based on Figure 5.10, the most common word in the non-depressive sentiment text is "day" with a frequency of 207. This suggests that the word "day" appears frequently in the text labeled as non-depressive sentiment. Other highly frequent words in the non-depressive sentiment text include "learn" (128), "life" (112), "spent" (109), and "find" (90).

The chart provides insights into the language patterns and prominent keywords that are prevalent in the text labeled as non-depressive sentiment. These words highlight positive experiences, personal growth, and enjoyment of life. By analyzing these common words, one can gain a better understanding of the themes and topics discussed in the text labeled as non-depressive sentiment.

5.2.6 Histogram of Text Lengths

Figure 5.11 below shows the Histogram of Text Lengths in the dataset.

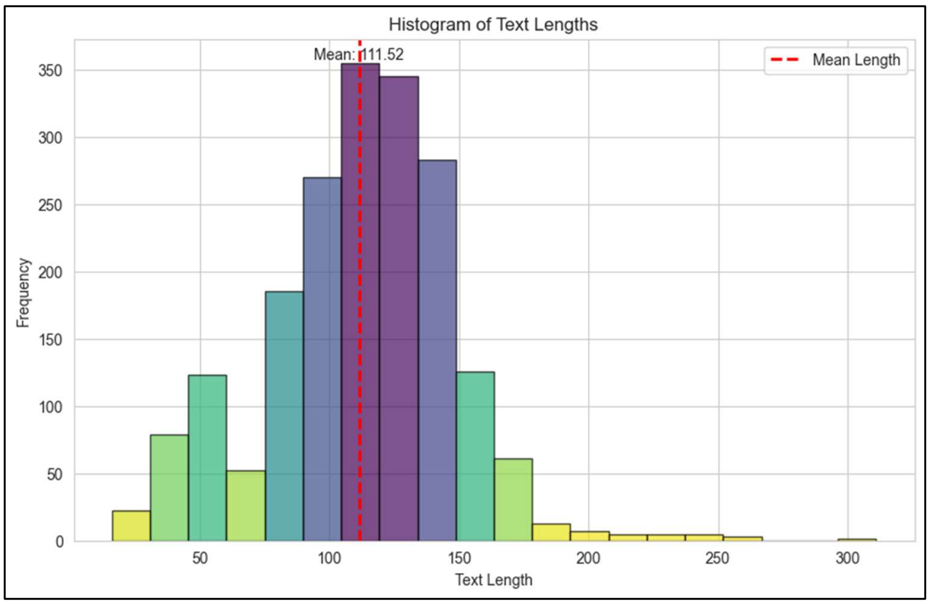


Figure 5. 11 Histogram of Text Lengths

Based on the Figure 5.11 above, the histogram of text lengths in the dataset exhibits a relatively symmetrical distribution. This means that there is a balance between shorter and longer text lengths, with a similar number of texts falling into both categories. In other words, there is no significant skew towards either extremely short or long texts. This balanced distribution implies that the texts in the dataset vary in length but do not tend to be excessively short or long. The mean text length, calculated to be 111.52, provides an average measure of the length of texts in the dataset. This value indicates that, on average, the texts consist of approximately 111 characters. Understanding the mean text length is valuable as it provides a representative estimate of the typical length of texts within the dataset. It serves as a useful reference point for understanding the overall quantity of text present in the dataset.

Taken together, the symmetrical distribution of text lengths and the average length of 111.52 suggest that the dataset comprises texts with a moderate length. The relatively balanced distribution indicates that the dataset encompasses texts of varying lengths, without a pronounced bias towards any range.

Here are some insights about the analyzation:

1. Balanced Text Lengths:
   The relatively symmetrical distribution suggests that there is no significant skew towards extremely short or long texts. This balanced distribution implies that the texts in the dataset vary in length but do not tend to be excessively short or long.
2. Moderate Text Length:
   The mean text length of 111.52 indicates that the texts are of moderate length. This information can be valuable for understanding the typical amount of text present in the dataset and can guide further analysis or modelling tasks.
3. Uniform Content:
   The symmetrical distribution and average text length provide insights into the nature of the dataset. A relatively uniform text length distribution might suggest that the texts come from similar sources or cover similar topics, resulting in a consistent length pattern.

5.3 Functionality Testing

Functionality testing is a type of testing that is focused on verifying that the software functions correctly and meets the specified requirements. The goal of functionality testing is to ensure that the software performs all the tasks it is designed to perform, and that it behaves as expected in different scenarios.

Functions in test cases are tested by entering input and evaluating output to verify whether the functions are successful, guaranteeing that the system's functionality operates as expected. Table 5.1 below shows the list of functionality test cases for this project.

Table 5. 1 List of Functionality Test Cases

Test Case	Expected Result	Success/ Failure
Responsive Design Testing	The website displays correctly, and functions as intended on different devices and screen sizes	Success
Dashboard Pages Testing	Users can view visualizations and data derived from the sentiment analysis model accurately and meaningfully	Success
Navigation Testing	All navigation links, buttons, and menus work correctly and guide users to the intended pages	Success
Anonymous Space Testing	Users can input text in the Anonymous Space page and receive sentiment analysis insights accurately	Success
Sentiment Analysis Testing	The sentiment analysis algorithm accurately analyses user input text and provides meaningful insights	Success
Articles Testing	Users can access and read articles smoothly, with all content displayed correctly	Success
Articles “What is Mental Health?”	Users can access and read articles “What is Mental Health?” smoothly, with all content displayed correctly	Success
Articles “What is Depression?”	Users can access and read articles “What is Depression?” smoothly, with all content displayed correctly	Success

Error Handling	Appropriate error messages are displayed for invalid inputs or actions	Success

5.4 Usability Testing

Usability testing is a type of testing that is focused on evaluating the user experience of a software application or product. Usability testing's objective is to guarantee that the target audience will find the software to be simple, clear, and effective.

To evaluate the usability and user satisfaction of my mental health website, a System Usability Testing (SUS) is conducted. This testing involved administering a series of carefully designed questions to participants, aiming to assess their perception of the website's ease of use, navigation, design, responsiveness, content relevance, and overall satisfaction.

The feedback obtained from this testing process will provide valuable insights into the user experience, highlight areas for improvement, and help us ensure that our website meets the needs and expectations of my target audience. Table 5.2 displays the Systems Usability Scale (SUS) Questions. The full form of the System Usability Scale (SUS) Questions can be view in the Appendices section in Appendix B.

Table 5. 2 System Usability Scale Questions

Number	**Overall Reaction to the System**	**Scale**				
1	On a scale of 1 to 5, how easy was it to use the website?	1	2	3	4	5
2	On a scale of 1 to 5, how intuitive was the website's navigation?	1	2	3	4	5
3	How satisfied are you with the overall design and layout of the website?	1	2	3	4	5
4	How would you rate the responsiveness and speed of the website?	1	2	3	4	5

5	Did you find it easy to understand the purpose and functionality of the website?	1	2	3	4	5
6	How well did the website meet your expectations in terms of content relevance and quality?	1	2	3	4	5
7	Did you encounter any difficulties or issues while using the website?	1	2	3	4	5
8	On a scale of 1 to 5, how likely are you to recommend this website to others?	1	2	3	4	5
9	Did the website provide a safe and secure environment for sharing thoughts and feelings?	1	2	3	4	5
10	Overall, how satisfied are you with your experience on the website?	1	2	3	4	5

The Table 5.3 below presents the results of the System Usability Testing (SUS) conducted with 20 respondents.

Table 5. 3 Results of the System Usability Testing

<table>
<tr><th>No</th><th>Results</th><th>Explanation</th></tr>
<tr><td>1.</td><td>

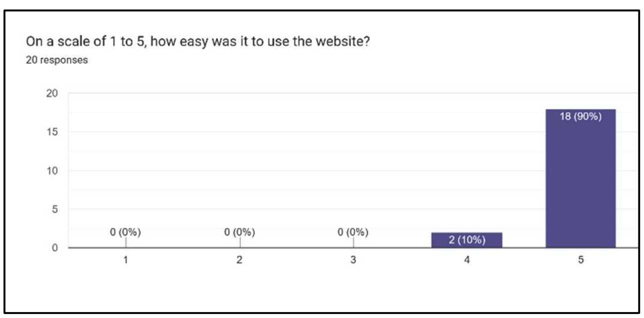
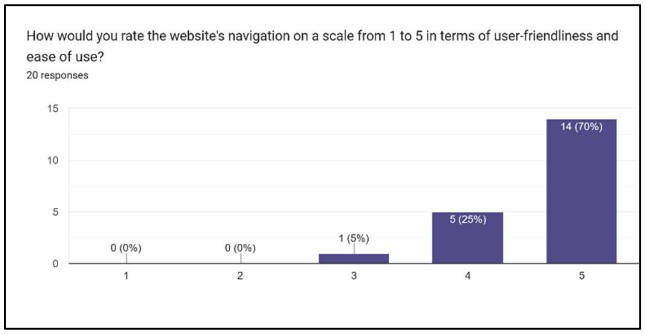
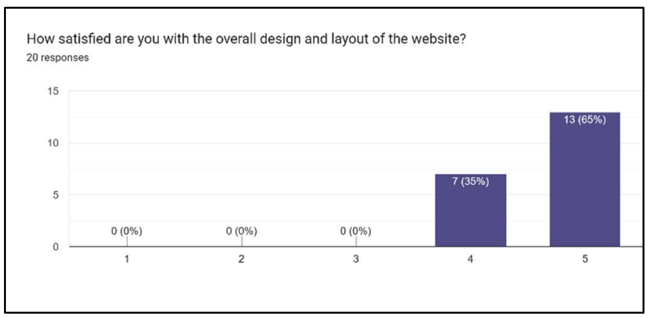
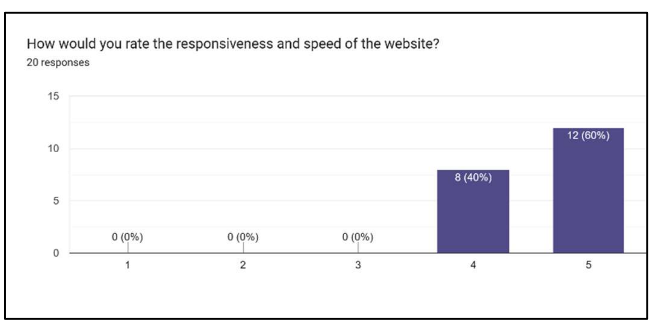
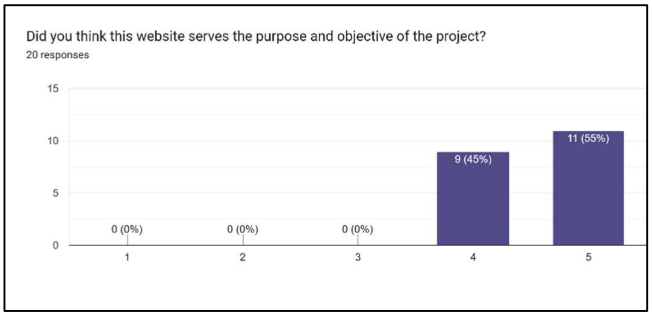
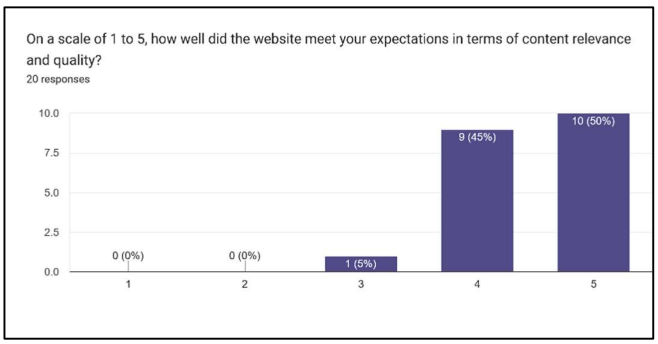
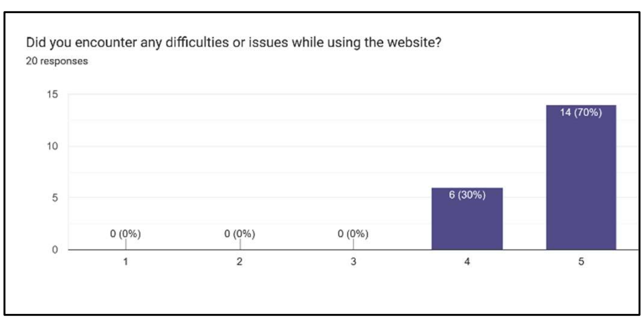


Figure 5. 12 Result SUS Question 1

</td><td>Based on Figure 5.12, 90% of the respondent choose the scale 5 and 10% choose scale 4.</td></tr>
</table>

2.	 **Figure 5. 13** Result SUS Question 2	Based on Figure 5.13, 70% of the respondent choose the scale 5, 25% choose scale 4 and 5% choose scale number 3.
3.	 **Figure 5. 14** Result SUS Question 3	Based on Figure 5.14, 65% of the respondent choose scale 5 and 35% choose scale 4.
4.	 **Figure 5. 15** Result SUS Question 4	Based on Figure 5.15, 60% choose scale 5 and 40% choose the scale 4.

5.	 **Figure 5. 16** Result SUS Question 5	Based on Figure 5.16, 55% choose the scale 5 and 45% choose the scale 4.
6.	 **Figure 5. 17** Result SUS Question 6	Based on Figure 5.17, 50% of the respondent choose scale 5 ,45% choose scale 4 and 5% choose the scale 3.
7.	 **Figure 5. 18** Result SUS Question 7	Based on Figure 5.18, 70% of the respondent choose scale 5 and 30% choose the scale 4.

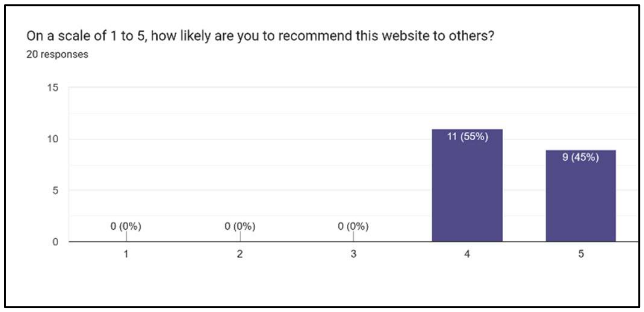
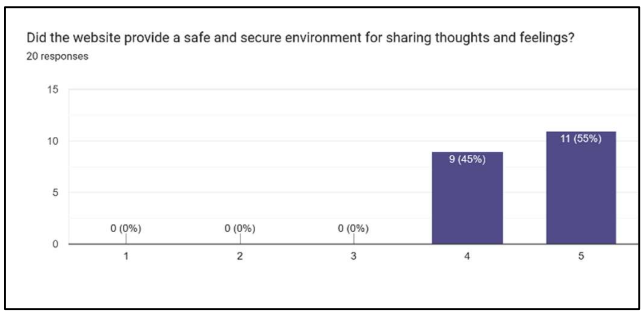
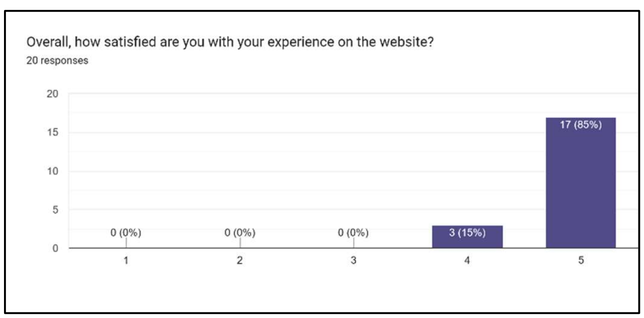
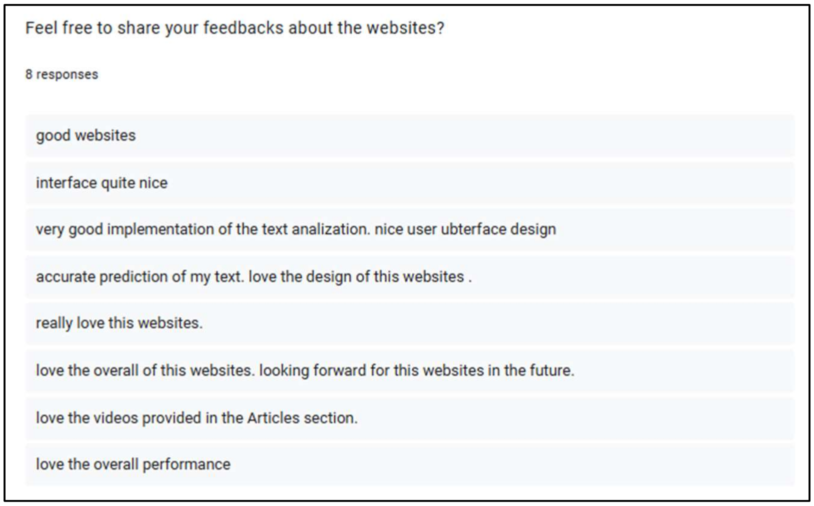
8.	 **Figure 5. 19** Result SUS Question 8	Based on Figure 5.19, 45% of the respondent choose scale 5 and 55% scale 4.
9.	 **Figure 5. 20** Result SUS Question 9	Based on Figure 5.20, 55% of the respondent choose scale 5 and 45% scale 4.
10.	 **Figure 5. 21** Result SUS Question 10	Based on Figure 5.21, 85% of the respondent choose scale 5 and 15% choose the scale 4.

Figure 5.22 below shows the respondent feedback gathered from 8 respondents about the websites.


Feel free to share your feedbacks about the websites?

8 responses

good websites

interface quite nice

very good implementation of the text analization. nice user ubterface design

accurate prediction of my text. love the design of this websites .

really love this websites.

love the overall of this websites. looking forward for this websites in the future.

love the videos provided in the Articles section.

love the overall performance

Figure 5. 22 Respondent Feedbacks

5.5 Accuracy Testing

The accuracy testing section serves as a vital assessment of the project's performance in accurately detecting depression through sentiment analysis. Two essential evaluation metrics, the Confusion Matrix using TF-IDF and the Classification Report, are employed to measure the effectiveness of the model. By analyzing these metrics, it can gain valuable insights into the model's predictive capabilities and its ability to distinguish between depressed and non-depressed sentiments. Through this comprehensive evaluation, we can ascertain the project's overall accuracy and its potential impact on mental health support.

5.5.1 Confusion Matrix using TF-IDF

Based on the provided accuracy testing below of the Confusion Matrix using TF-IDF, the model achieved an impressive accuracy percentage of 98.97%. This indicates that the model performed exceptionally well in classifying the data points correctly. Figure 5.23 below shows the accuracy percentage with Confusion Matrix using TF-IDF.

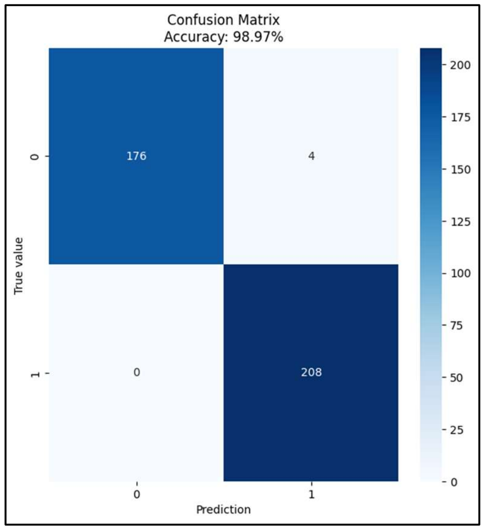


Figure 5. 23 TF-IDF Confusion Matrix

Now let's analyze the individual components of the Confusion Matrix:

TN (True Negative): The model correctly predicted 176 instances as non-depressive texts.

FP (False Positive): The model incorrectly predicted 4 instances as depressive texts when they were actually non-depressive. This means there were 4 instances where the model had a false alarm or a false positive prediction.

FN (False Negative): The model did not make any false negative predictions, meaning it correctly classified all the depressive texts without misclassifying them as non-depressive.

TP (True Positive): The model correctly predicted 208 instances as depressive texts.

Overall, the Confusion Matrix indicates that the model had a high level of accuracy and performed well in both detecting non-depressive texts (TN) and depressive texts (TP). The low values for FP and FN suggest that the model had a minimal number of misclassifications, which is a positive outcome.

5.5.2 Classification Report

Figure 5.24 below provided the classification report provides a detailed evaluation of the model's performance on each class, along with the overall metrics.

Classification Report:

	precision	recall	f1-score	support
0	1.00	0.98	0.99	180
1	0.98	1.00	0.99	208
accuracy			0.99	388
macro avg	0.99	0.99	0.99	388
weighted avg	0.99	0.99	0.99	388

Figure 5. 24 Classification Report

1. Precision:

   Precision measures the accuracy of positive predictions. For class 0 (non-depressive text), the precision is 1.00, indicating that all the instances predicted as non-depressive were actually non-depressive. For class 1 (depressive text), the precision is 0.98, indicating that the model correctly classified 98% of the instances as depressive.

2. Recall:

   Recall measures the ability of the model to correctly identify the positive class. For class 0, the recall is 0.98, meaning the model successfully captured 98% of the actual non-depressive instances. For class 1, the recall is 1.00, indicating that the model correctly identified all the depressive instances.

3. F1-score:

   The F1-score is the harmonic mean of precision and recall, providing a balanced measure of the model's performance. For both classes, the F1-score is 0.99, indicating a high level of accuracy and balance between precision and recall.

4. Support:

Support represents the number of instances in each class. In this case, there are 180 instances of class 0 and 208 instances of class 1.

The overall accuracy of the model is 0.99, which aligns with the accuracy percentage mentioned earlier. The macro average and weighted average of precision, recall, and F1-score are also 0.99, indicating consistent performance across both classes.

Overall, the classification report confirms that the model achieved high accuracy and performed well in classifying both non-depressive and depressive texts. It provides a comprehensive evaluation of the model's performance on each class, allowing for a detailed understanding of its strengths in detecting depressive texts.

CHAPTER 6

CONCLUSION AND RECOMMENDATION

In this chapter, this project's outcomes and future suggestions are presented. The strengths observed during the project's development will be highlighted, followed by the identification of limitations to understand the project's scope better. Additionally, key recommendations will be provided to enhance the platform's effectiveness in the future. Finally, significant findings and outcomes will be discussed, emphasizing the importance of the mental health web platform with sentiment analysis in detecting depression. These insights pave the way for further advancements, making a positive impact on mental health support and awareness.

6.1 Project Strengths

Here's some revised version the strengths of this project.

1. Purpose and Impact:
   This project addresses the critical issue of mental health by providing a secure online platform where users can express their thoughts and feelings. By incorporating sentiment analysis, this project aims to detect depressive sentiments and provide insights and resources to individuals who may be experiencing mental health challenges. This project's focus on mental health and well-being showcases its potential to make a positive impact on individuals' lives.
2. Data Collection and Analysis:
   This project has collected a diverse dataset from multiple sources, including Twitter API and studies by reputable researchers. This comprehensive dataset allows for an in-depth analysis of user-generated content and expert-validated data related to depression symptoms. By leveraging sentiment analysis algorithms and machine

learning techniques, the project can accurately classify and identify depressive sentiments, enabling early detection and support.

3. Data Visualization:
   This project utilizes various visualizations, such as word clouds, bar charts, and pie charts, to present sentiment distributions, common words, and patterns within the dataset. These visualizations provide an intuitive and accessible way for users and stakeholders to interpret and understand the data, enabling them to gain valuable insights and make informed decisions.
4. User-Focused Features: This project includes user-centric features like the Anonymous Space and Sentiment Analysis pages, providing users with a safe and anonymous platform to share their thoughts and receive sentiment analysis insights. Additionally, the Mental Health Resources and Articles pages offer educational content and support, enhancing the overall user experience and providing valuable resources for mental health awareness and well-being.
5. Technical Implementation: This project showcases proficiency in implementing machine learning algorithms, natural language processing techniques, and web development frameworks. The use of widely used libraries like Scikit-learn, pandas, and nltk demonstrates the project's ability to leverage existing tools and technologies effectively.

The strengths of this project lie in its purposeful focus on mental health, comprehensive data analysis, and user-friendly features. These strengths position the project as a valuable tool for detecting and addressing depression symptoms, while providing support and resources for individuals in need.

6.2 Project Limitations

The limitations of this project should be considered to provide a comprehensive understanding of its scope and potential constraints. Here are the limitations that has been figured in this project.

1. Data Bias and Generalizability:
   The data that have been collected for this project comes from sources like Twitter, which means it may not represent everyone's experiences. People on social media can have different backgrounds and behaviours, so the findings from the sentiment analysis may not apply to everyone equally. It needs to be cautious when drawing conclusions from this data.
2. Subjectivity of Sentiment Analysis:
   The sentiment analysis algorithms are designed to understand emotions in text, but it's not always easy to capture the full range of human emotions and experiences through words alone. Sometimes people express themselves in ways that can be hard to interpret, leading to potential inaccuracies in the sentiment analysis results.
3. Reliance on Self-Reported Data:
   This project only relies on users to input their own text, which means we're dependent on how accurately they express their feelings. People may have different ways of describing their emotions, and there's a chance of biases or misinterpretations in their self-reported data. This could affect the reliability of the results.
4. Limited Scope of Sentiment Analysis:
   While sentiment analysis gives insights into the emotional tone of text, it can't capture the full complexity of mental health conditions like depression. Depression involves various symptoms and factors beyond sentiment alone. Our sentiment analysis model may not be able to detect subtle signs of depression, limiting its effectiveness in identifying individuals who may need professional help.

6.3 Project Recommendations

In looking ahead to the future of this project, it is important to consider certain recommendations that can further enhance its scope and potential impact. Acknowledging the project's limitations, several areas have been identified where future improvements can be made:

1. Expand Data Collection:
   Consider collecting data from diverse sources beyond social media platforms to capture a broader range of experiences and perspectives. This could involve collaborating with mental health organizations, conducting surveys, or integrating data from other online communities.
2. Improve Sentiment Analysis Algorithms:
   Continuously refine and enhance the sentiment analysis algorithms used in the project. Explore advanced natural language processing techniques, deep learning models, or ensemble methods to improve accuracy and capture more nuanced emotional expressions.
3. Collaborate with Mental Health Professionals:
   Partner with mental health professionals, psychologists, and experts in the field to validate and enhance the effectiveness of your sentiment analysis model. Their insights and expertise can help refine the model to better identify potential signs of depression and provide appropriate support and resources.
4. Implement Real-Time Monitoring: Develop a system that can monitor sentiment and emotional trends in real-time. This can help identify emerging patterns, detect potential mental health crises, and provide timely interventions or resources to individuals in need.

6.4 Project Conclusion

In conclusion, I am delighted to state that all the objectives of this project have been successfully achieved. The mental health website has been designed to provide a safe and anonymous space for users to freely express their thoughts and feelings. The sentiment-analysis algorithm, developed using Support Vector Machine (SVM), effectively determines whether a user is exhibiting symptoms of depression based on their input text. Notably, the accuracy testing results have yielded an impressive 98.97% accuracy, showcasing the reliability and precision of the website system. Rigorous testing has further ensured the functionality and usability of the platform, contributing to its successful implementation. This project's accomplishments represent a significant stride towards addressing mental health concerns and providing valuable support to users in their emotional well-being journey.

REFERENCES

Akhtar, P., Ma, L., Waqas, A., Naveed, S., Li, Y., Rahman, A., & Wang, Y. (2020). Prevalence of depression among university students in low and middle income countries (LMICs): a systematic review and meta-analysis. *Journal of Affective Disorders*, *274*(May), 911–919. https://doi.org/10.1016/j.jad.2020.03.183

Alanazi, S. A., Khaliq, A., Ahmad, F., Alshammari, N., Hussain, I., Zia, M. A., Alruwaili, M., Rayan, A., Alsayat, A., & Afsar, S. (2022). Public's Mental Health Monitoring via Sentimental Analysis of Financial Text Using Machine Learning Techniques. *International Journal of Environmental Research and Public Health*, *19*(15). https://doi.org/10.3390/ijerph19159695

Ashraful Islam, M., Yun Low, W., Ting Tong, W., Wan Yuen, C. C., & Abdullah, A. (2018). Factors Associated with Depression among University Students in Malaysia: A Cross-sectional Study. *KnE Life Sciences*, *4*(4), 415. https://doi.org/10.18502/kls.v4i4.2302

Cardona, M. A. (2021). Supporting Child and Student Social, Emotional, Behavioral, and Mental Health Needs. *U.S. Department of Education*, 1–99. https://www2.ed.gov/documents/students/supporting-child-student-social-emotional-behavioral-mental-health.pdf

Cassady, J. C., Pierson, E. E., & Starling, J. M. (2019). Predicting Student Depression With Measures of General and Academic Anxieties. *Frontiers in Education*, *4*(February), 1–9. https://doi.org/10.3389/feduc.2019.00011

Chaudhari, A. R., & Joshi, S. D. (2021). Study of effect of Agile software development Methodology on Software Development Process. *Proceedings of the 2nd International Conference on Electronics and Sustainable Communication Systems, ICESC 2021*, 1848–1851. https://doi.org/10.1109/ICESC51422.2021.9532842

Chiong, R., Budhi, G. S., Dhakal, S., & Chiong, F. (2021). A textual-based featuring approach for depression detection using machine learning classifiers and social media texts. *Computers in Biology and Medicine*, *135*(February), 104499. https://doi.org/10.1016/j.compbiomed.2021.104499

Croock, J., Mpinganjira, M. G., Gathoo, K., Bulmer, R., Lautenberg, S., Dlamini, Q., Londani, P., Solontsi, A., Stevens, C., & Francis, J. M. (2023). Probable depression and its correlates among undergraduate students in Johannesburg, South Africa. *Frontiers in Psychiatry*, *14*(February), 1–8. https://doi.org/10.3389/fpsyt.2023.1018197

Gao, L., Xie, Y., Jia, C., & Wang, W. (2020). Prevalence of depression among Chinese university students: a systematic review and meta-analysis. *Scientific Reports*, *10*(1), 1–11. https://doi.org/10.1038/s41598-020-72998-1

Gibbon, S., Khalifa, N. R., Cheung, N. H. Y., Völlm, B. A., & McCarthy, L. (2020). Psychological interventions for antisocial personality disorder. *Cochrane Database of Systematic Reviews*, *2020*(9). https://doi.org/10.1002/14651858.CD007668.pub3

GmbH, A. (2020). *Front End (In a Website)*. Airfocus. https://airfocus.com/glossary/what-is-a-front-end/

Guo, W. (2022). Applications of Logistic Regression and Naive Bayes in Commodity Sentiment Analysis. *ACM International Conference Proceeding Series*, 224–230. https://doi.org/10.1145/3531232.3531265

Gupta, P., Gupta, M., Kansal, R., Singh, L., Khotani, M., & Gupta, V. S. (2021). Analysis of Twitter data to Identify Intuitive Mental Health Well Being. *2021 5th International Conference on Information Systems and Computer Networks, ISCON 2021*, 2021–2024. https://doi.org/10.1109/ISCON52037.2021.9702319

Hamza, C. A., Ewing, L., Heath, N. L., & Goldstein, A. L. (2021). When social isolation is nothing new: A longitudinal study on psychological distress during COVID-19 among university students with and without preexisting mental health concerns. *Canadian Psychology*, *62*(1), 20–30. https://doi.org/10.1037/cap0000255

Herdiansyah, H., Roestam, R., Kuhon, R., & Santoso, A. S. (2023). Their post tell the truth: Detecting social media users mental health issues with sentiment analysis. *Procedia Computer Science*, *216*(2022), 691–697. https://doi.org/10.1016/j.procs.2022.12.185

Hussna, A. U., Trisha, I. I., Ritun, I. J., & Alam, M. G. R. (2021). COVID-19 impact on students' Mental Health: Explainable AI and Classifiers. *2021 International Conference on Decision Aid Sciences and Application, DASA 2021*, 847–851. https://doi.org/10.1109/DASA53625.2021.9682371

Keshavarzi, F., Megan, C., Sin, C. K., Wei, C. C., And, C. C. M., & Yin, C. W. Y. (2021). *Research Article the Impact of Covid-19 Pandemic on the Mental Health of College and University*. *22*(October).

Kokkinos, C. M., Tsouloupas, C. N., & Voulgaridou, I. (2022). The effects of perceived psychological, educational, and financial impact of COVID-19 pandemic on Greek university students' satisfaction with life through Mental Health. *Journal of Affective Disorders*, *300*(September 2021), 289–295. https://doi.org/10.1016/j.jad.2021.12.114

Leal Filho, W., Wall, T., Rayman-Bacchus, L., Mifsud, M., Pritchard, D. J., Lovren, V. O., Farinha, C., Petrovic, D. S., & Balogun, A. L. (2021). Impacts of COVID-19 and social isolation on academic staff and students at universities: a cross-sectional study. *BMC Public Health*, *21*(1), 1–19. https://doi.org/10.1186/s12889-021-11040-z

McGowan, M. L., Norris, A. H., & Bessett, D. (2020). Care Churn — Why Keeping Clinic Doors Open Isn't Enough to Ensure Access to Abortion. *New England Journal of Medicine*, *383*(6), 508–510. https://doi.org/10.1056/nejmp2013466

Medhat, W., Hassan, A., & Korashy, H. (2017). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011

Prommas, P., Lwin, K. S., Chen, Y. C., Hyakutake, M., Ghaznavi, C., Sakamoto, H., Miyata, H., & Nomura, S. (2023). The impact of social isolation from COVID-19-related public health measures on cognitive function and mental health among older adults: A systematic review and meta-analysis. *Ageing Research Reviews*, *85*(December 2022), 101839. https://doi.org/10.1016/j.arr.2022.101839

Rahman, R. A., Haziqah Mohamad Zaini, F., Nizam Mohd Danuri, M. S., & Amin, A. (2022). The Sentiment Analysis on Mental Health Awareness by Non-Governmental Organisation's Twitter. *IVIT 2022 - Proceedings of 1st International Visualization, Informatics and Technology Conference*, 185–190. https://doi.org/10.1109/IVIT55443.2022.10033345

Reavley, N., Fernando, L. M. N., & Jorm, A. (2021). Websites and the validity of mental health care information. In *Mental Health in a Digital World*. Elsevier Inc. https://doi.org/10.1016/B978-0-12-822201-0.00009-5

Rohith Gandhi. (2018). *Support Vector Machine — Introduction to Machine Learning Algorithms*. Towards Data Science. https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

Sadasivuni, S. T., & Zhang, Y. (2019). Analyzing tweets to discover twitter users' mental health status by a word-frequency method. *Proceedings - 2019 IEEE International Conference on Intelligent Systems and Green Technology, ICISGT*

2019, *2*, 5–8. https://doi.org/10.1109/ICISGT44072.2019.00016

Salza, P., Musmarra, P., & Ferrucci, F. (2019). Agile Methodologies in Education: A Review. *Agile and Lean Concepts for Teaching and Learning*, 25–45. https://doi.org/10.1007/978-981-13-2751-3_2

Sanci, L., Kauer, S., Thuraisingam, S., Davidson, S., Duncan, A. M., Chondros, P., Mihalopoulos, C., & Buhagiar, K. (2019). Effectiveness of a mental health service navigation website (link) for young adults: Randomized controlled trial. *JMIR Mental Health*, *6*(10), 1–17. https://doi.org/10.2196/13189

Seidel, E. J., Mohlman, J., Basch, C. H., Fera, J., Cosgrove, A., & Ethan, D. (2020). Communicating Mental Health Support to College Students During COVID-19: An Exploration of Website Messaging. *Journal of Community Health*, *45*(6), 1259–1262. https://doi.org/10.1007/s10900-020-00905-w

Shokrzade, A., Ramezani, M., Akhlaghian Tab, F., & Abdulla Mohammad, M. (2021). A novel extreme learning machine based kNN classification method for dealing with big data. *Expert Systems with Applications*, *183*(May), 115293. https://doi.org/10.1016/j.eswa.2021.115293

Stuart, H. (2016). Reducing the stigma of mental illness. *Lancet*, *357*(9262), 1055. https://doi.org/10.1016/S0140-6736(00)04269-0

Subramani, Chellamuthu. Subramanian, K. (2017). Academic Stress and Mental Health Among High School Students. *Indian Journal of Applied Research*, *7*(5), 405. https://doi.org/10.5923/j.ijpbs.20150501.04

Subramani, C., & Kadhiravan, S. (2017). Academic stress and mental health among high school students. *Indian Journal Of Applied Research*, *7*(5), 404–406. https://www.researchgate.net/profile/Subramani_Chellamuthu4/publication/321572693_Academic_Stress_and_Mental_Health_Among_High_School_Students/links/5a277809a6fdcc8e866e6419/Academic-Stress-and-Mental-Health-

Among-High-School-Students.pdf?_sg%5B0%5D=hbNbA

Sutherland, P. L. (2018). the Impact of Mental Health Issues on Academic Achievement in High School Students. *Electronic Theses, Projects, and Dissertations*. https://scholarworks.lib.csusb.edu/etd/660

Termonia, B. (2021). *Applied Text Mining and Sentiment Analysis with Python*. Udemy. https://www.udemy.com/course/applied-text-mining-and-sentiment-analysis-with-python/learn/lecture/24697916#overview

Trotzek, M., Koitka, S., & Friedrich, C. M. (2020). Utilizing Neural Networks and Linguistic Metadata for Early Detection of Depression Indications in Text Sequences. *IEEE Transactions on Knowledge and Data Engineering*, *32*(3), 588–601. https://doi.org/10.1109/TKDE.2018.2885515

Uddin, M. Z., Dysthe, K. K., Følstad, A., & Brandtzaeg, P. B. (2022). Deep learning for prediction of depressive symptoms in a large textual dataset. *Neural Computing and Applications*, *34*(1), 721–744. https://doi.org/10.1007/s00521-021-06426-4

Villarroel, M. A., & Terlizzi, E. P. (2020). Symptoms of Depression Among Adults: United States, 2019. *NCHS Data Brief*, *379*, 1–8.

World Health Organization. (2022). *Mental disorders*. WHO. https://www.who.int/news-room/fact-sheets/detail/mental-disorders

Xu, Q. A., Chang, V., & Jayne, C. (2022). A systematic review of social media-based sentiment analysis: Emerging trends and challenges. *Decision Analytics Journal*, *3*(April), 100073. https://doi.org/10.1016/j.dajour.2022.100073

Yadav, S. (2022). Detecting Presence of Mental Illness Using Nlp Sentiment Analysis. *International Research Journal of Modernization in Engineering Technology and Science*, *06*, 2582–5208.

https://www.irjmets.com/uploadedfiles/paper/issue_6_june_2022/27031/final/fin_irjmets1656403378.pdf

Yap, S. Y., Foo, C. N., Lim, Y. M., Ng, F. L., Mohd Sidik, S., Tang, P. Y., Najar Singh, J. K., & Pheh, K. S. (2021). *Prevalence of Depression and its Associated Risk Factors among University Students in Malaysia: A Pilot Study*. 9005. https://doi.org/10.3390/ecerph-3-09005

Yue, L., Chen, W., Li, X., Zuo, W., & Yin, M. (2019). A survey of sentiment analysis in social media. *Knowledge and Information Systems*, *60*(2), 617–663. https://doi.org/10.1007/s10115-018-1236-4

APPENDICES

APPENDIX A: GANTT CHART

TASK			MONTH / WEEK
Task ID	Task Description	Start Date	End Date
1	Identify the area and the purpose of the proposal system	11/10/2022	14/10/2022
2	Set the project title	17/10/2022	21/10/2022
3	Selection of supervisor	24/20/22	28/20/22
4	Define the problem statements	31/10/2022	4/11/2022
5	Identify the objectives, scope, and significance of the project	7/11/2022	11/11/2022
7	Write literature review	14/11/2022	9/12/2022
8	Identify project methodology	12/12/2022	16/12/2022
9	Design flowchart, use case and user interface.	19/12/2022	23/12/2022
10	Develop the Sentiment Analysis Model using SVM	20/3/2023	19/4/2023
11	Build the system prototype	22/2/2023	11/6/2023
12	Conduct system testing, including the functionality and usability testing.	14/6/2023	17/7/2023

APPENDIX B: SYSTEM USABILITY SCALE (SUS) QUESTIONS

System Usability Scale (SUS)

Assalamualaikum and hello,

My name is Muhammad Aiman Naim bin Mohd Faizul and I am a Final Year student studying Bachelor of Computer Science (Hons.) from College of Computing, Informatics and Mathematics at UiTM Melaka, Jasin Campus. My Final Year Project (FYP) requires me to assess my system's usability on the topic mentioned above.

This survey is an initiative form to gather information about the system usability scale and it is aimed specifically for people who have been selected to participate in the usability testing phases. All the data gathered in this survey is for research purposes only and will be kept confidential. Your cooperation is highly appreciated.

Thank you again for your participation.

2020490132@student.uitm.edu.my Switch account

Not shared

* Indicates required question

Name *

Your answer

On a scale of 1 to 5, how easy was it to use the website? *

1	2	3	4	5
○	○	○	○	○

How would you rate the website's navigation on a scale from 1 to 5 in terms of user-friendliness and ease of use? *

1	2	3	4	5
○	○	○	○	○

How satisfied are you with the overall design and layout of the website? *

1	2	3	4	5
○	○	○	○	○

How would you rate the responsiveness and speed of the website? *

1	2	3	4	5
○	○	○	○	○

Did you think this website serves the purpose and objective of the project? *

	1	2	3	4	5	
No	○	○	○	○	○	Yes

On a scale of 1 to 5, how well did the website meet your expectations in terms of content relevance and quality? *

1	2	3	4	5
○	○	○	○	○

Did you encounter any difficulties or issues while using the website? *

	1	2	3	4	5	
Yes	○	○	○	○	○	No

On a scale of 1 to 5, how likely are you to recommend this website to others? *

1	2	3	4	5
○	○	○	○	○

Did the website provide a safe and secure environment for sharing thoughts and feelings? *

1	2	3	4	5
○	○	○	○	○

Overall, how satisfied are you with your experience on the website? *

1	2	3	4	5
○	○	○	○	○

Feel free to share your feedbacks about the websites?

Your answer

Submit

Clear form

APPENDIX C: LEAN CANVAS MODEL

Designed for (Product name): A SENTIMENT ANALYSIS APPROACH TO UNIVERSITY MENTAL HEALTH		**Designed by:** **Name and student ID:** MUHAMMAD AIMAN NAIM BIN MOHD FAIZUL : 2020490132	**Approved by:** **Supervisor name, stamp and signature**	**Date:** 20/4/2023
PROBLEM	**SOLUTION**	**UNIQUE VALUE PROPOSITION**	**UNFAIR ADVANTAGE**	**CUSTOMER SEGMENTS**
• Misinformation about mental health online is a problem that needs to be addressed. • People with mental health issues need a safe space to express themselves. • There is a need for more resources and investment in mental health.	• The sentiment analysis model will provide insights into the emotional experiences and attitudes of university students towards mental health, helping to identify individuals who may be in need of support and resources. • The website interface will provide university students with easy access to information and resources related to mental health, addressing the issue of misinformation and the lack of resources and investment in mental health issues.	• Because it has the ability to provide accessible mental health resources and real-time sentiment analysis of social media data related to mental health among university students.	• Only focusing university student.	• University student who are seeking information and resources related to mental health.
REVENUE STREAM	**KEY METRICS**	**COST STRUCTURE**	**CHANNELS**	**EARLY ADOPTERS**
• Donations • Advertising revenue	• Sentiment analysis accuracy • Engagement rate	• Development costs • Hosting costs	• Instagram • Twitter	• University students who are interested in mental health and wellness, and who are open to using technology to support their mental health.

APPENDIX D: APPROVAL LETTER

Universiti Teknologi MARA
KM 26 Jalan Lendu
78000 Alor Gajah
Melaka Bandaraya Bersejarah
Tel: +606 558 2000
Faks: +606 558 2001

UNIVERSITI TEKNOLOGI MARA

Our reference : 600CM(PJI/RMU.5/5/12)
Date : 14th April 2023

MUHAMMAD AIMAN NAIM BIN MOHD FAIZUL
Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA, Cawangan Melaka,
Kampus Jasin,
77300 Melaka.

Dear Sir,

APPROVAL LETTER - UiTM MELAKA RESEARCH ETHICS

The Branch Ethics Review Committee (BERC) UiTM Cawangan Melaka has considered and approved your Research Ethics application.

Details of the approval are as follows:

Referral No.	BERC/MLK/117/2023
Proposal Title	Mental Health Websites For University Students With Sentiment Analysis
Approval Period	11th October 2022 - 13th August 2023
Authorised personnel	Madam Azida Binti Mohamed Noh

Condition/s of Approval

- Research must be conducted according to the approved proposal.
- The submission of the final report must be submitted to the Ethics Office on or before the anniversary of approval and completion of the subject.
- You must report as soon as practicable anything that might warrant a review of ethical approval of the project, including:
  - Serious or unexpected adverse events (which should be reported within 72 hours)
  - Unforeseen events that might affect the ethical acceptability of the project.
- Any changes to the proposal must be approved prior to their implementation (except where an amendment is undertaken to eliminate immediate risk to participants)

Yours sincerely,

DR. NUR HAYATI ABD RAHMAN
Chairperson
UiTM Branch Research Ethics Committee

c.c.: Rector, UiTM Melaka Branch