

ESTIMATING DEPRESSIVE TENDENCIES
OF TWITTER USER VIA SOCIAL MEDIA
DATA

LOH HOOI TENG

Bachelor of Computer Science
(Computer Systems & Networking)
With Honours

UNIVERSITI MALAYSIA PAHANG

UNIVERSITI MALAYSIA PAHANG

DECLARATION OF THESIS AND COPYRIGHT

Author's Full Name : LOH HOOI TENG

Date of Birth :

Title : ESTIMATING DEPRESSIVE TENDENCIES OF TWITTER
USER VIA SOCIAL MEDIA DATA

Academic Session : 2022/2023

I declare that this thesis is classified as:

- CONFIDENTIAL (Contains confidential information under the Official Secret Act 1997)*
- RESTRICTED (Contains restricted information as specified by the organization where research was done)*
- OPEN ACCESS I agree that my thesis to be published as online open access (Full Text)

I acknowledge that Universiti Malaysia Pahang reserves the following rights:

1. The Thesis is the Property of Universiti Malaysia Pahang
2. The Library of Universiti Malaysia Pahang has the right to make copies of the thesis for the purpose of research only.
3. The Library has the right to make copies of the thesis for academic exchange.

Certified by:

(Student's Signature)

(Supervisor's Signature)

Dr. Nur Shazwani binti Kamarudin

New IC/Passport Number
Date: 17 January 2023

Name of Supervisor
Date: 17 January 2023

NOTE : * If the thesis is CONFIDENTIAL or RESTRICTED, please attach a thesis declaration letter.

THESIS DECLARATION LETTER (OPTIONAL)

Librarian,
Perpustakaan Universiti Malaysia Pahang,
Universiti Malaysia Pahang,
Lebuhraya Tun Razak,
26300, Gambang, Kuantan.

Dear Sir,

CLASSIFICATION OF THESIS AS RESTRICTED

Please be informed that the following thesis is classified as RESTRICTED for a period of three (3) years from the date of this letter. The reasons for this classification are as listed below.

Author's Name
Thesis Title

Reasons (i)

(ii)

(iii)

Thank you.

Yours faithfully,



(Supervisor's Signature)

Date: 17 January 2023

Stamp: DR. NUR SHAZWANI KAMARUDIN
PENSYARAH KANAN
UNIVERSITI MALAYSIA PAHANG
26600 PEKAN, PAHANG.
TEL : 09-424 4736

Note: This letter should be written by the supervisor, addressed to the Librarian, *Perpustakaan Universiti Malaysia Pahang* with its copy attached to the thesis.



SUPERVISOR'S DECLARATION

I hereby declare that I have checked this thesis and in my opinion, this thesis is adequate in terms of scope and quality for the award of the degree of Bachelor of Computer Science in Computer Systems and Networking.

A handwritten signature in black ink, appearing to read 'Nur Shazwani', is written above a horizontal line.

(Supervisor's Signature)

Full Name : Dr. Nur Shazwani binti Kamarudin

Position : Senior Lecturer

Date : 17 January 2023

(Co-supervisor's Signature)

Full Name :

Position :

Date :



STUDENT'S DECLARATION

I hereby declare that the work in this thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at Universiti Malaysia Pahang or any other institutions.

Hooi Teng

(Student's Signature)

Full Name : LOH HOOI TENG

ID Number : CA19072

Date : 17 January 2023

ESTIMATING DEPRESSIVE TENDENCIES OF TWITTER USER VIA SOCIAL
MEDIA DATA

LOH HOOI TENG

Thesis submitted in fulfillment of the requirements
for the award of the degree of
Bachelor of Computer Science in Computer Systems and Networking

Faculty of Computer Systems and Software Engineering

UNIVERSITI MALAYSIA PAHANG

JANUARY 2023

ACKNOWLEDGEMENTS

Alhamdulillah, I'm relieved and thankful that I'll be able to finish this assignment or research. I'd want to convey my heartfelt thanks to Universiti Malaysia Pahang for providing me with the opportunity to perform my final year project and hone the skills I've already learned.

Furthermore, I would like to express my gratitude to my closest supervisor, Dr. Nur Shazwani binti Kamarudin, for her patience, tolerance, and guidance in suffering all of my challenges till I completed my research.

I would want to express my heartfelt thanks and appreciation to my family and friends for their unwavering support in pushing me forward and cheering me up anytime I felt sad.

Last but not least, I would want to express my gratitude to everyone who helped me out with hands, support, and time, both directly and indirectly.

ABSTRAK

Pada masa kini, kemurungan adalah masalah kesihatan mental utama yang memberi kesan kepada semua peringkat umur, jantina dan etnik di seluruh dunia. Orang ramai berasa semakin selesa berkongsi idea mereka di laman web atau aplikasi media sosial secara praktikal setiap hari dalam era komunikasi dan teknologi kontemporari ini. Matlamat projek ini adalah untuk mengkaji sifat-sifat teks yang berkaitan dengan kecenderungan kemurungan melalui dataset Twitter. Dalam projek ini memerlukan set data yang besar daripada twitter, jadi kami akan mengumpul set data Twitter daripada tapak web Kaggle yang sudah mempunyai set data twitter lengkap yang boleh dimuat turun untuk melaksanakan anggaran kecenderungan kemurungan pengguna Twitter. Dataset twitter boleh digunakan untuk menguji tahap kecenderungan kemurungan dengan tiga algoritma pembelajaran mesin yang berbeza. Tiga algoritma pembelajaran mesin yang berbeza ini iaitu Support Vector Machine, XGBoost, dan Random Forest. Kami akan menggunakan tiga algoritma pembelajaran mesin ini untuk membandingkan ketepatan dan prestasi pengguna twitter yang tertekan. Oleh itu, pembelajaran mesin yang berbeza mempunyai jenis ciri yang berbeza yang boleh digunakan untuk menjalankan anggaran kecenderungan kemurungan.

ABSTRACT

Nowadays, depression is a major mental health problem that affects people of all ages, genders, and ethnicities all over the world. People feel increasingly comfortable sharing their ideas on social media websites or applications practically every day in this age of contemporary communication and technology. The aim of this project is to study the properties of the text related to depressive tendencies via Twitter dataset. In this project will require huge dataset from twitter, so we will collect the Twitter dataset from Kaggle websites that already have the completed twitter dataset that can be downloaded in order to implement the estimating depressive tendencies of Twitter user. The twitter dataset can be used to test the level of depressive tendencies with three different machine learning algorithms. These three different machine learning algorithms which are Support Vector Machine, XGBoost, and Random Forest. We will use these three machine learning algorithms to compare the accuracy and performance of the depressed twitter user. Therefore, different machine learning have different types of features that can use to conduct the estimating depressive tendencies.

TABLE OF CONTENT

DECLARATION	
TITLE PAGE	
ACKNOWLEDGEMENTS	ii
ABSTRAK	iii
ABSTRACT	iv
TABLE OF CONTENT	v
LIST OF TABLES	viii
LIST OF FIGURES	ix
LIST OF ABBREVIATIONS	xi
CHAPTER 1 INTRODUCTION	12
1.1 Introduction	12
1.2 Problem Statements	14
1.3 Objectives	15
1.4 Scope	15
1.5 Report Organization	16
CHAPTER 2 LITERATURE REVIEW	17
2.1 Introduction	17
2.2 Prior Research Work	17
2.3 Summary	21
CHAPTER 3 METHODOLOGY	22
3.1 Introduction	22

3.2	Research Framework	23
3.2.1	Data Collection/Extraction	23
3.2.2	Data pre-processing	23
3.2.3	Modelling	24
3.2.4	Evaluate and Result	24
3.3	Project Requirement	25
3.3.1	Input	25
3.3.2	Output	25
3.3.3	Process Description	29
3.3.4	Constraints And Limitations	30
3.4	Proposed Design	31
3.4.1	Flowchart	31
3.4.2	Explanation of the Flowchart	32
3.5	Data Design	33
3.5.1	Dataset Description	33
3.6	Proof Of Initial Concept	34
3.6.1	Machine Learning Algorithm	34
3.6.2	Python Machine Learning Libraries	37
3.7	Software Equipment	39
3.8	Hardware Equipment	40
3.9	Potential Use Of Proposed Solution	41
	CHAPTER 4 RESULTS AND DISCUSSION	42
4.1	Introduction	42
4.2	Result	42
4.3	Discussion	53

CHAPTER 5 CONCLUSION	54
5.1 Conclusion	54
5.2 Limitation	55
5.3 Future Work	55
REFERENCES	56
APPENDIX A	58

LIST OF TABLES

Table 2.1	Compare between prior studies	21
Table 3.1	Software equipment and its specification	40
Table 3.2	Hardware equipment and its specification	41
Table 4.1	Accuracy, Precision, Recall, and F1-Score Measure on Twitter Depression Dataset Using Different ML Classifiers on Training and Testing	43

LIST OF FIGURES

Figure 3.1	Research Framework	23
Figure 3.2	Example for Accuracy Formula	26
Figure 3.3	Example for Precision Formula	27
Figure 3.4	Example for Recall Formula	27
Figure 3.5	Example for F1 Score Formula	28
Figure 3.6	Example of Confusion Matrix	29
Figure 3.7	Flowchart	31
Figure 3.8	Twitter Dataset from Kaggle	33
Figure 3.9	Example of Support Vector Machine	34
Figure 3.10	Example of XGBoost Classifier	35
Figure 3.11	Formula of XGBoost	36
Figure 3.12	Example of Random Forest	37
Figure 3.13	Example of Numpy Libraries in command	38
Figure 3.14	Example of Pandas Libraries in command	38
Figure 3.15	Example of NLTK Libraries in command	39
Figure 3.16	Example of Matplotlib Libraries in command	39
Figure 4.1	Accuracy, Precision, Recall, and F1-Score Measure on Twitter Depression Dataset Using Different ML Classifiers on Training and Testing	44
Figure 4.2	Result of Accuracy Measure for Different Classifiers in both Training and Testing	45
Figure 4.3	Result of Precision Measure for Different Classifiers in both Training and Testing	46
Figure 4.4	Result of Recall Measure for Different Classifiers in both Training and Testing	46
Figure 4.5	Result of F1-Score Measure for Different Classifiers in both Training and Testing	47
Figure 4.6	All common words in the entire dataset	48
Figure 4.7	Positive words which is non-depressive tweets in the dataset	48
Figure 4.8	Negative words which is depressive tweets in the dataset	49
Figure 4.9	Hyperparameters in XGBoost machine learning classifier	50
Figure 4.10	Hyperparameters in Random Forest machine learning classifier	50
Figure 4.11	Confusion Matrix for Training with Random Forest classifier	51
Figure 4.12	Confusion Matrix for Testing with Random Forest classifier	51
Figure 4.13	Estimating Depressive Tendencies on Tweets	52

Figure 4.14	Result is 1 (Depressive Tendencies) after inserting the tweet “I am depress”	52
Figure 4.15	Result is 0 (Non-Depressive Tendencies) after inserting the tweet “I am happy”	53

LIST OF ABBREVIATIONS

SVM	Support Vector Machine
XGBoost	Extreme Gradient Boosting
RF	Random Forest
NLTK	Natural Language Toolkit
NLP	Natural Language Processing
NB	Naïve Bayes
DT	Decision Tree
KNN	K-Nearest Neighbour

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

Recently, social media has become a big part in our daily life. It allows people to communicate with their pals more easily. Images, videos, and Tweets they publish, or watch are entirely dependent on their mental health, idea generation, knowledge growth, and current events in our society. They also let us express our feelings. With the introduction of difference, it is now simple to construct social media sites that convey our thoughts and feelings (Neelavathi et al., 2021). But when we use social media too much like twitter, it affects the mood of users when they see extreme comments on the site, it makes others tend to get depressed and fall into it, which causes them to become more and more inactive thinking, so that suicidal thoughts emerge. This is because depression has recently emerged as one of the most urgent problems facing the globe. According to a survey done by the World Health Organization (WHO) in 2012, 350 million individuals globally suffer from depression (Tsugawa et al., 2013). World Health Organization stated the great majority of depressed people never seek therapy. Because depressed people tend to be sensitive and oblivious of their own state, the observed depression is challenging to assess (Sawangarreerak & Thanathamatee, 2020). This is especially concerning for the younger generation, who are more likely to blame others and have low self-esteem before seeking treatment. Depression is frequently undetected and ignored, even during visits to a primary health care practitioner (Nadeem, 2016). Because depression may lead to suicide, the rise in depressed persons is seen as a severe issue (Tsugawa et al., 2013).

It is critical to recognize a person's depression tendency to provide successful therapy. According to the WHO, the majority of individuals who require depression therapy do not receive it. Depression is not often detected and appropriately diagnosed,

and persons with depression are frequently undertreated, according to primary care clinicians. As a result, it is critical to detect that the patient or those around him are prone to depression in order to effectively treat depression (Tsugawa et al., 2013).

The socioeconomic status of an individual may be negatively impacted by depression (Aleem S, Huda N, 2022). Depression makes people more reluctant to socialize. Depression can influence anyone at any moment in their lives, and it can have terrible implications, the worst-case scenario being suicide. As a result, it's critical to swiftly and precisely detect the signs and symptoms of depression based on their tweets (Azam et al., 2021). Therefore, we can use the sentiment analysis through machine learning techniques to identify whether the people who tweets is prone to depressive tendencies.

1.2 PROBLEM STATEMENTS

Nowadays, social media coverage of depressive-prone events is associated with depressive behaviours. Concerns have been raised about this issue. This is due to the fact that depression has become the most popular mental health condition of interest among computational social scientists. It is a very common mental illness that impacts a wide range of behaviours and communication patterns. According to a recent analysis of a significant metropolitan area, over half (45%) of all cases of severe depression were incorrectly identified, demonstrating that depression underdiagnosis is still a problem (Reece et al., 2017). To determine if someone has depressed tendencies or thoughts in their daily activities, we can no longer depend simply on health records, demographic data, clinical interviews, or patient self-report questionnaires. With changes in technology and the rise of media communication, users may develop depression tendencies. The most commonly used social network today is Twitter, as it is the most important area for collecting data to detect depressive tendencies.

Sentiment analysis is one of the methods that emerged with social networks in automatic language processing. Especially on social media platforms, we need to use more adequate techniques to conduct impromptu research on the status quo. This is because on the Twitter social network, it is difficult to analyze and detect how many people are prone to depression. We should use machine learning algorithms to process large amounts of unstructured data for sentiment analysis to help detect depressive tendencies in tweets. By using sentiment analysis, we can analyze whether the user speech in the data on Twitter contains a melancholic tendency.

1.3 OBJECTIVE

This project has three objectives, which are as follows:

- To study the properties of the text related to depressive tendencies via Twitter dataset
- To leverage the machine learning algorithm for dataset of the twitter user related to depressive tendencies
- To test the accuracy of the depressive tendencies of twitter user with machine learning techniques for estimating the depressive tendencies

1.4 SCOPE

The scopes are listed as follows:

- i. The study of the use of dataset via social media data in twitter platform
- ii. The target user which is the researcher or the admin person who want to classify the data from the related word of depression
- iii. The dataset will be classified by using machine learning algorithm.
- iv. Dataset will be divided into training (70%) and testing (30%).
- v. The result will show the accuracy of the depressive tendencies of twitter user

1.5 REPORT ORGANIZATION

The thesis consists of five (5) chapters. The introduction, problem background, objectives, scope, and thesis organization are all included in the first chapter, Chapter 1 Introduction.

The second chapter is the Chapter 2 Literature Review which includes the literature review or background of study, and critical review of comparison. The comparison of approaches, methods, hardware, technology, and past studies, as well as analyze the advantages and disadvantages will also be covered in this chapter.

The third chapter is the Chapter 3 Methodology which consists of the project planning methods, hardware or software used. This chapter explains the structure, development approach, and processes to get the desired output.

The fourth chapter is the Chapter 4 Implementation, Result, and Testing. This chapter which discusses the results acquired via the use of method and technique and will go through the project's outcomes in depth.

The last chapter is the Chapter 5 Conclusion which includes the overall project requirements and discusses the project development prototype's limitations as well as future work.

CHAPTER 2

LITERATURE REVIEW

2.1 INTRODUCTION

This chapter consists of literature review for prior research work that related with the title of estimating depression on Twitter user. In this chapter will compare the previous existing research studies about the domain that be apply in the research project.

2.2 PRIOR RESEARCH WORK

This study (P. Kumar et al., 2022) proposed feature based depression detection from twitter data using machine learning techniques by using the Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), and XGBoost classifier, and data will be collected from Twitter users via Twitter APIs to train the different algorithm for depression detection. By using the Twitter APIs, two datasets will be created which are depression and non-depressed users. For a period of one month beginning with an anchor tweet, a straightforward set of statistics will be gathered for each Twitter user's profile and tweets. Number was used and average per tweet was calculated for the hashtags, links, replies, and @mentions. A tweet plus the attendant replies and comments make up a social media session. Instead of combining all of a user's tweets, each tweet is regarded as a post for a session in our work. The proposed effort seeks to analyse as few tweets as possible, analyse each tweet separately, and decide if a person is depressed or not. Four machine learning models were proposed on a specified feature set to determine the depressive individuals. The different types of the linguistic feature were used to extract the features from the tweets that can assist in depression detection. The different machine learning algorithm were also implemented in order to perform the best accuracy for depressed tweet classification. The Nature Language Processing (NLP) was also be executed in this project. In accordance with this finding, the linguistic features (Trigram + TF - IDF) AND LDA were evaluated the best

performance for depression detection with the accuracy of 89% via the Support Vector Machine Classifier.

This project (Priya et al., 2020) use five different machine learning algorithm on raw employed and unemployed participant response data to predict psychological problems of depression. In this study, data was collected by using the standard questionnaire and measure with Depression, Anxiety and Stress Scale questionnaire (DASS-21) from 348 participants aged between 20 until 60 years old through Google forms. After the completion of data collection, the replies of the participants were encoded using numeric values ranging from 0 to 3, and the scores were determined by summing the values associated with each question and using the method which are score equal to the sum of the rating point of each class multiple with 2. After the final scores were determined, they were categorized into Normal, Mild, Moderate, Severe, and Extremely Severe categories. The machine learning methods were implemented using RStudio version 3.5 and the R programming language. This forecasts the percentage of persons who suffer from stress, anxiety, or depression, based on the degree of their symptoms. The training and test sets were split in a 70:30 ratio, reflecting the training and test sets, respectively. After that, classified by using the five different machine learning algorithm which are Decision Tree (DT), Random Forest (RF), Naïve Bayes, Support Vector Machine (SVM), and K-Nearest Neighbour (KNN). After five different machine algorithm was implemented in five different results, it will show the accuracy precision, and f1 score for each class in five different algorithms. To test the accuracy of the machine learning algorithm, it refers to the accuracy and f1 score, respectively. Based on the result, the accuracy of the Naïve Bayes was the most accurate and highest result compare with another 4 machine learning algorithms.

This project (H. S. Alsagri & Ykhlef, 2020) proposed the machine learning-based approach for depression detection in Twitter using content and activity features. The tasks of data preparation, feature extraction, and classification are carried out in this study using a variety of R packages and the RStudio IDE. The classifiers are trained using 10-fold cross-validation and then assessed on a held-out test set in order to reduce overfitting. Depression will be identified using the activity and content features (DDACF) classification model. The classification model also includes TF-IDF. For data collection,

it has been collected from 500 users with more than 1M tweets and for 334 users were classified as depressed. These tweets are reviewed by a human annotator to ensure that users are speaking about their own depression and not that of a friend or family member. Non-depressed people are selected at random and carefully reviewed to verify that no tweets containing the character string "depress" are posted. Users with less than five Twitter postings are eliminated in order to reduce noisy and inaccurate data. A variety of classification algorithms, including SVM with various kernels (Linear and Radial), DT, and NB, will be used in this research. According to the findings, the SVM model has achieved the best accuracy metric combinations because it transforms a wildly non-linear classification problem into one that can be divided linearly.

Comparison between the prior research work

Elements	Research 1	Research 2	Research 3
Author	Feature Based Depression Detection from Twitter Data Using Machine Learning Techniques (P. Kumar et al., 2022)	Predicting Anxiety, Depression and Stress in Modern Life using Machine Learning Algorithms (Priya et al., 2020)	Machine Learning-Based Approach for Depression Detection in Twitter Using Content and Activity Features (H. S. Alsagri & Ykhlef, 2020)
Domain	To evaluate Twitter user postings and determine characteristics that could point to depressed tendencies in internet users.	To identify the best accuracy of the five different machine learning algorithms.	To employ machine learning techniques to locate a potential depressed Twitter user based on the individual's tweets and network behavior.

Technique	Natural Language Processing (NLP), techniques by using Logistic Regression, Support Vector Machine (SVM), Random Forest, and XGBoost algorithm based on linguistic features	Decision Tree (DT), Random Forest (RF), Naïve Bayes, Support Vector Machine (SVM), and K-Nearest Neighbour (KNN)	Machine learning classified with Support Vector Machine (SVM), Naïve Bayes (NB), and Decision Tree (DT)
Data	Collected data from Twitter users via Twitter APIs	Collected data from 348 participants via Google forms using standard questionnaire and through DASS-21, the Depression, Anxiety and Stress Scale questionnaire	Collected data from 111 user profiles in Twitter
Advantages	The SVM and XGBoost classifiers perform well on the SA measure, with the SVM classifier having the maximum detection accuracy by combining features.	Naïve Bayes is most accurate when using five different machine learning algorithms to test the accuracy.	The amount of words in the corpus is reduced via SVM-linear, which reduces calculation time.
Disadvantages	Because of comparison between different machine learning classifiers based on different linguistic features, it needed	Because of the problem of unbalanced classes, the best-model selection was based on the f1 score.	SVM requires a long training period on huge datasets.

	some time to do the coding process.		
Limitation	The XGBoost and SVM classifiers show the best performance on different features, so it hard to identify which is the best machine learning algorithm.	The limitation on comparing results is needed to compare the accuracy and f1-score to determine the most accurate machine learning algorithm.	-

Table 2.1 Compare between prior studies

2.3 SUMMARY

Based on the previous research work, every research studies have their methods and approaches, advantages, and disadvantages for the machine learning algorithm. In summary, the most efficient for research that I prefer is the Research 3 which is Machine Learning-Based Approach for Depression Detection in Twitter Using Content and Activity Features because it uses the effective methods which is the SVM-linear classifier and achieved the optimal accuracy of detecting depression in Twitter and there is no limitation on these studies.

CHAPTER 3

METHODOLOGY

3.1 INTRODUCTION

This chapter will discuss the process on estimating the depressive tendencies by using the method and technique from raw Twitter dataset. The machine learning will be use in this research project to fulfill the objective as mentioned before. Artificial intelligence (AI) in the form of "machine learning" (ML) enables programmes to make predictions more accurately without having explicit coding. Machine learning algorithms will use historical data as input and forecast future output values. To estimate the best result of accuracy for the machine learning, it needs the requirement and the methods for estimating depressive tendencies from raw twitter data. This project will use the Support Vector Machine (SVM), XGBoost, and Random Forest (RF) because need to compare between these three machine learning algorithms to find the best accuracy for estimating depression.

3.2 RESEARCH FRAMEWORK

In this research framework can explore the study approach for using the raw Twitter dataset to estimate the depressive tendencies.

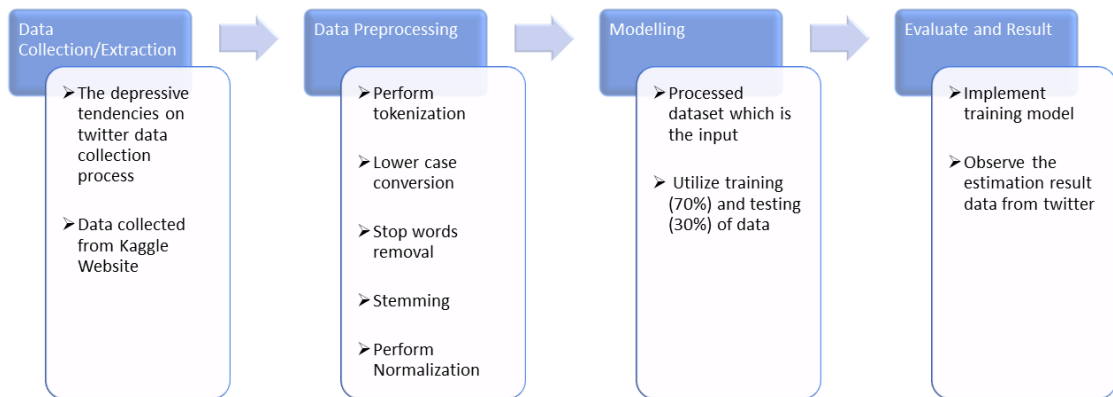


Figure 3.1 Research Framework

3.2.1 Data Collection/Extraction

This project will use the Twitter dataset about estimating depressive tendencies from the Kaggle website to find as in Kaggle have a lot of the dataset from different topic that involved. Besides that, it is more convenient to use it because it is easy to get and download the dataset from this website. For Twitter dataset that get from Kaggle website is the previous user commented or tweets in the Twitter and has been collected to become a dataset gather to become a depression dataset.

3.2.2 Data pre-processing

NLTK (Natural Language Toolkit) was used to analyze the data collected from the Kaggle website. This is because the text data in Twitter dataset will mainly all in text format which is in strings. To complete the task, a machine learning algorithm requires a numerical feature vector. All text data will be transformed to lowercase since the algorithm does not recognise the same words in different case as separate. Tokenization, lower case conversion, stop words removal, stemming, and other operations may all be

carried out via NLTK. Moreover, tokenization will also be used to explain the conversion of regular text strings into a list of tokens that include the necessary words. For more information, to discover a list of sentences, a sentence tokenizer is utilized (Asad et al., 2019). NLTK can also convert all the text data into lowercase and remove all the stop words such as he, she, the, her, and so on before processing for a clean model. All the stop words can be viewed in the English language via NLTK. Through stemming in NLTK, when the word like helping, helped, helpful have the root word and express same meaning, so root word can be extracted and discard the remainder. NLTK can also take off all the stop words, @ mentions, and retweets (Singh & Choudhary, 2017). Noise is also eliminated such as the anything that isn't in normal numbers or characters. Additionally, normalisation is employed to remove punctuation, retweets, mentions, links, unknown emoji, and symbols from text (H. S. Alsagri & Ykhlef, 2020).

3.2.3 Modelling

In modelling, dataset will separate into the ratio 70:30 of training and testing which are 70% and 30%, respectively. The input and expected output are included in the training data, as well as the input and corresponding expected output. For testing data, 30% of testing data will use to assess whether the training model is doing well on the unseen data. Machine learning tools like Support Vector Machine (SVM), XGBoost, and Random Forest (RF) are used to create classifications.

3.2.4 Evaluate and Result

The training model will choose three machine learning algorithms which are SVM, XGBoost, and RF to estimate the depressive tendencies of Twitter user by comparing three different methods of algorithms. The machine learning algorithm will be used to implement the estimate experiment with training and test data. For training process, it will define the best accuracy result in comparing three machine learning models. After comparing the models, the result will display which one is the most consistent with the dataset. For the result of estimating depressive tendencies will display in charts and diagrams.

3.3 PROJECT REQUIREMENT

In this section will mainly talk about the input, output, process description, constraints, and limitations.

3.3.1 INPUT

In project requirement, the input for the estimating depressive tendencies of Twitter user data is the processed dataset that download from the Kaggle website. For twitter dataset that downloaded from Kaggle is .csv file, which inside have all twitter user data and is uploaded 2 years ago. This twitter dataset is related with the depressive tendencies of twitter user via social media data. We can use that twitter dataset to perform the task of data pre-processing and classification in order to determine the accuracy of the result of estimating depressive tendencies of twitter user.

3.3.2 OUTPUT

For output of this project is the expected output for estimating depressive tendencies of Twitter user which is the estimating result of the accuracy by using the three different machine learning algorithms which are Support Vector Machine (SVM), XGBoost, and Random Forest (RF). The result will show the accuracy, precision, recall, f1 score and so on to support the evidence that the best machine learning algorithm in estimating depressive tendencies in Twitter user data.

Accuracy

When it comes to accuracy, the proportion of correctly predicted observations to all observations is the simplest performance statistic to comprehend. Considering our model's accuracy, one would assume that it is the best. Yes, accuracy is a great indication. However, this is only true when the dataset's false positive and false negative rate values are almost equal. Therefore, in order to evaluate the efficacy of our model, we must take into account additional elements.

To calculate the accuracy of the machine learning algorithm, below shows that the formula for accuracy:

$$Acc = \frac{truePositive + trueNegative}{truePositive + trueNegative + falsePositive + falseNegative}$$

		Predicted class	
		Class = Yes	Class = No
Actual Class	Class = Yes	True Positive	False Negative
	Class = No	False Positive	True Negative

Figure 3.2 Example for Accuracy Formula

For more information:

True Positive (TP) = These are the positive values that are correctly predicted, which means that the actual class has a value of yes and the predicted class has a value of yes.

True Negative (TN) = These are the negative values that are correctly predicted, which means that the actual class has a value of no and the predicted class has a value of no.

False Positive (FP) = When the actual class contradicts the predicted class, which means that the actual class has a value of no and the predicted class has a value of yes.

False Negative (FN) = When the actual class contradicts the predicted class, which means that the actual class has a value of yes and the predicted class has a value of no.

Precision

Precision can be calculated using either of the classes. Accuracy of the negative class refers to the classifier's capacity to refuse to categorise a negative sample as positive. Accuracy of the positive class refers to a classifier's ability to avoid categorising a positive sample as negative. The maximum and minimum values for precision are 1 and 0, respectively.

To calculate the precision of the machine learning algorithm, below shows that the formula for precision:

$$P = \frac{\textit{truePositives}}{\textit{truePositives} + \textit{falsePositives}}$$

Figure 3.3 Example for Precision Formula

Recall

Recall measures the proportion of correctly anticipated positive results to all instances in the actual class. The machine learning model is better at identifying positive from negative situations when the recall score is higher.

To calculate the recall of the machine learning algorithm, below shows that the formula for recall:

$$R = \frac{\textit{truePositives}}{\textit{truePositives} + \textit{falseNegatives}}$$

Figure 3.4 Example for Recall Formula

F1 Score

Precision and Recall are weighted averaged to provide the F1 score. Therefore, while computing this score, erroneous positives as well as false negatives are taken into account. F1 is often more beneficial than accuracy, especially if you have an unequal class distribution, despite the fact that it is less intuitively understandable than accuracy. Accuracy works best when the cost of false positives and false negatives is about equal. If there is a large difference in the costs of false positives and false negatives, it is desirable to incorporate both Precision and Recall.

To calculate the f1 score of the machine learning algorithm, below shows that the formula for f1 score:

$$\text{F1 Score} = \frac{2 * (\text{Precision} * \text{Recall})}{\text{Precision} + \text{Recall}}$$

Figure 3.5 Example for F1 Score Formula

Confusion Matrix

A confusion matrix is a table that shows how many accurate and erroneous classifications a classifier produced. It is used to assess the performance of a classification model. It may be used to measure a classification model's success by computing performance metrics including accuracy, precision, recall, and F1-score (Li et al., 2021). The target variable can take on one of two positive or negative values. The rows display the anticipated values for the target variable, while the columns display its actual values. The True Positives (TP), which signify situations in which the actual value and anticipated value are both positive, offer extra information about the confusion matrix. In the case of True Negatives (TN), both the forecast and the actual value are negative. False positives (FP), commonly referred to as Type 1 errors, occur when the true value is negative, but the forecast is positive. False Negatives (FN), which are often referred to as

Type 2 mistakes, are instances where the actual outcome is positive but the forecast was negative.

		ACTUAL VALUES	
		Positive	Negative
PREDICTED VALUES	Positive	TP	FP
	Negative	FN	TN

Figure 3.6 Example of Confusion Matrix

3.3.3 PROCESS DESCRIPTION

In this project, the process will have five stage which are the data collection or extraction, data pre-processing, classification, analyze and result. For the first process is the data collection which is also called data extraction, which mainly collected the dataset from the Twitter data that find in the Kaggle website. After that, the second process is the data pre-processing which is using the NLTK after collected the data. The NLTK will mainly discuss the text processing which are tokenization, lower case conversion, remove stop words, stemming and so on. The third procedure is modelling, which divides the dataset primarily into the proportions of 70% for training and 30% for testing, respectively. For machine learning algorithm will mainly use the three algorithms which are Support Vector Machine (SVM), XGBoost, and Random Forest (RF) to perform the result of estimating depressive tendencies of each method. The last process is the evaluation and result of the estimating depressive tendencies based on the different measures in different machine learning algorithms such as accuracy, precision, f1 score and so on.

3.3.4 CONSTRAINTS AND LIMITATIONS

In this project, the constraint for estimating depressive tendencies from Twitter data is this project is only using three machine learning which are Support Vector Machine (SVM), XGBoost, and Random Forest (RF). Moreover, it will become a constraint as not using other machine learning algorithm which determine the best machine learning algorithm. Besides that, for different types of machine learning have many complicated processes to develop and propose, so it may need some time to completely understand the process of each machine learning algorithm undergo.

For the limitation of this project is needed to use both result of accuracy and f1 score for three machine learning algorithms to determine which is the best machine learning algorithm in this project. This is because if only using the accuracy result to determine the best methods, it will become inaccurate, so it must also include the f1 score to estimate the best methods for three different machine learning algorithms to estimate depressive tendencies of Twitter datasets. For another limitation is the language limitations will be eliminated, and several languages will be regarded as samples.

3.4 PROPOSED DESIGN

3.4.1 Flowchart

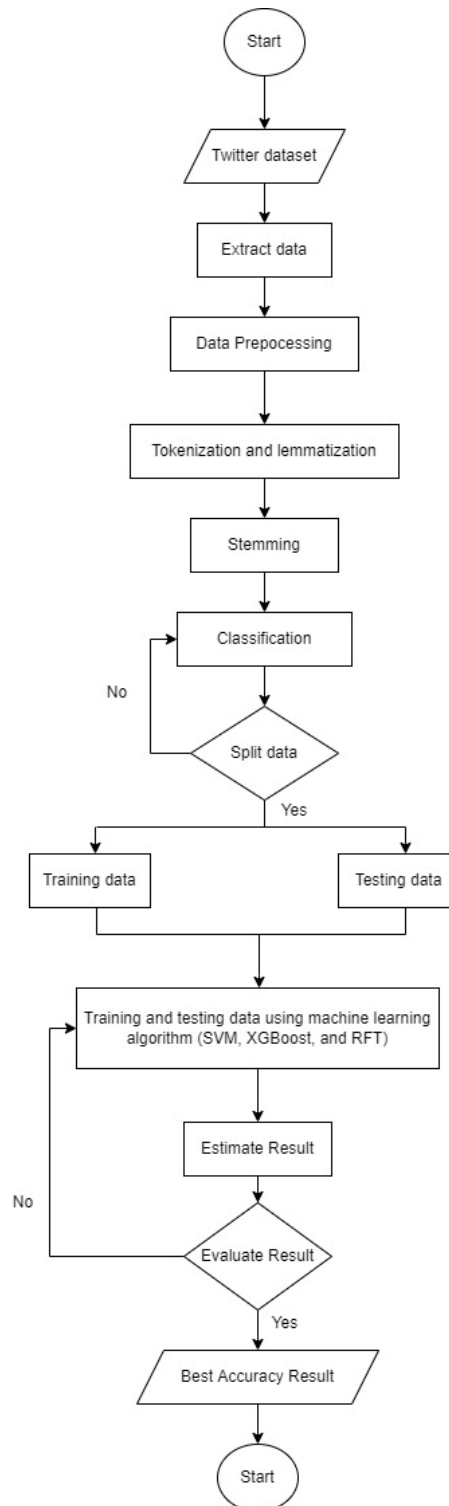


Figure 3.7 Flowchart

3.4.2 Explanation of the Flowchart

The process will begin by extracting data. Data extraction is the process of gathering or getting various sorts of data from a variety of sources. Data will be acquired from the Kaggle website that related with the estimating depressive tendencies of twitter user via social media data and has been extracted to become a dataset for this study. Therefore, twitter dataset that find in Kaggle website will be used in this project to estimate depressive tendencies of twitter user.

After that, data pre-processing will be implemented to remove the stop words, @ mentions, URLs, retweets and eliminate anything that is not in normal numbers or character. Because the algorithm does not distinguish between the same words written in different case-insensitive ways, the data must be converted to lowercase as part of the pre-processing step. Tokenization and lemmatization will be conducted in data pre-processing. Tokenization refers to the process of turning regular text strings into a list of tokens, which are the terms needed for and associated with this project. For lemmatization, it normally aims to eliminate the only word endings and restore a word to its base or dictionary form. Moreover, stemming process will also be carried out in this process (Hassan et al., 2017).

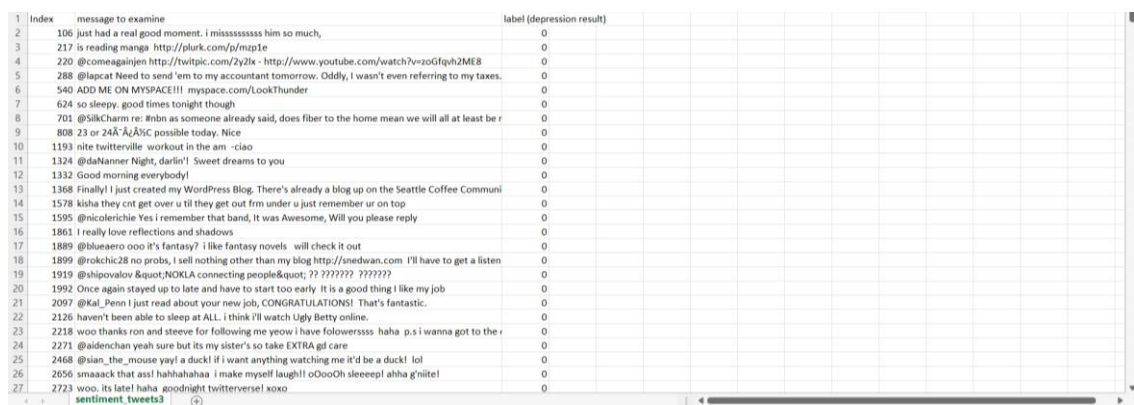
Moreover, the dataset that get from Kaggle website will be divided into two parts which are testing and training. For training data is used to execute the three different machine learning algorithms and for the testing data, the trained model is applied to the supervised dataset and uses the model to link the training and testing datasets to produce the estimation's outcome.

Therefore, the data will be estimated, and the result will be given for estimating depressive tendencies in twitter user. It will define the best accuracy result in different three machine learning algorithm models.

3.5 DATA DESIGN

3.5.1 Dataset Description

In this project, the twitter dataset which is downloaded from the Kaggle website, <https://www.kaggle.com/datasets/gargmanas/sentimental-analysis-for-tweets/code> named as Sentimental Analysis for Tweets, which depressed and non-depressed tweets were gathered from twitter. In this dataset will have around 10314 of tweets. This dataset is in .csv file and was uploaded 2 years ago and is collected from Twitter APIs. This twitter dataset will need to download to perform the task of data pre-processing about the data cleaning process and classification to estimate the depressive tendencies of twitter user. For this Twitter dataset have some features which are the Index, message to examine, and label (depression result). Moreover, this dataset is not clean yet. Therefore, this dataset will do the data cleaning, data pre-processing, and so on to classify whether is prone to depression or not.



Index	message to examine	label (depression result)
106	just had a real good moment. i missssssss him so much,	0
217	is reading manga http://plurk.com/f/mxp1e	0
220	@comeagainjen http://twitpic.com/2y2lk - http://www.youtube.com/watch?v=zoGfqh2ME8	0
288	@lapcat Need to send 'em to my accountant tomorrow. Oddly, i wasn't even referring to my taxes.	0
540	ADD ME ON MYSPEACE!!! myspace.com/LookThunder	0
624	so sleepy. good times tonight though	0
701	@SilkCharm re: #nbn as someone already said, does fiber to the home mean we will all at least be r	0
808	23 or 24 Å Å Å Å Å possible today. Nice	0
1193	nite twiterville workout in the am -clao	0
1324	@daNanner Night, darlin'! Sweet dreams to you	0
1332	Good morning everybody!	0
1368	Finally! I just created my WordPress Blog. There's already a blog up on the Seattle Coffee Communi	0
1578	kisha they cnt get over u til they get out frm under u just remember ur on top	0
1595	@nicolerichie Yes i remember that band, It was Awesome, Will you please reply	0
1861	I really love reflections and shadows	0
1889	@blueaero ooo it's fantasy? i like fantasy novels will check it out	0
1899	@rokchic28 no probs, I sell nothing other than my blog http://smedwan.com i'll have to get a listen	0
1919	@shipovator "NOKIA connecting people" ?? ?????? ??????	0
1992	Once again stayed up to late and have to start too early. It is a good thing i like my job	0
2097	@Kal_Penn i just read about your new job, CONGRATULATIONS! That's fantastic.	0
2126	haven't been able to sleep at ALL. i think i'll watch Ugly Betty online.	0
2218	woo thanks ron and steeve for following me yeow i have folowerssss haha .p.s i wanna got to the r	0
2271	@aidenchan yeah sure but its my sister's so take EXTRA gd care	0
2468	@sian_the_mouse yay! a duck! if i want anything watching me it'd be a duck! lol	0
2656	smaaack that ass! hahahaha i make myself laugh!! oDooOh sleeep! ahha g'nite!	0
2723	woo. its late! haha goodnight (twitterverse) xoxo	0

Figure 3.8 Twitter Dataset from Kaggle

3.6 PROOF OF INITIAL CONCEPT

3.6.1 Machine Learning Algorithm

3.6.1.1 Support Vector Machine

Support Vector Machine (SVM), a machine learning approach, is most frequently used for classification issues but may also be utilised for regression difficulties (Priya et al., 2020). This classifier has lately been employed in several applications because of its exceptional classification capabilities and presentation quality. It creates two distinct classes from the data, often referred to as hyperplanes, with the greatest distance between them. SVM produces a number of different hyperplanes and will choose the one that best separates the data points (Azam et al., 2021). It can fine-tune several characteristics while maintaining outstanding performance to avoid overfitting.

An SVM function called a kernel assists in issue resolution. They provide solutions for challenging calculations. The good thing about kernel is that it makes it possible to move fluidly over higher dimensions while doing calculations. It can grow to an infinite number of dimensions by using kernels (Kamde et al., 2022). The SVM's training process produces a latent hyperplane that divides the examples into two groups. The distance between the segmented hyperplane and the nearby training samples increases as a result. Both the side of the hyperplane that the item is leaning on and the side of the hyperplane that the object is leaning on may be predicted by the SVM (H. S. Alsagri & Ykhlef, 2020).

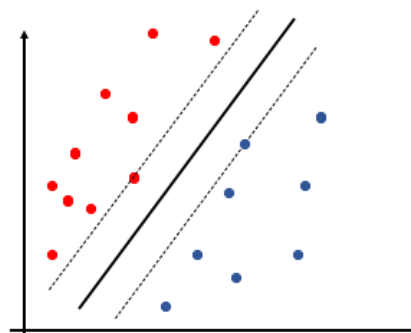


Figure 3.9 Example of Support Vector Machine

3.6.1.2 XGBoost (Extreme Gradient Boosting)

The gradient-boosted decision tree (GBDT) machine learning toolkit XGBoost, which stands for "Extreme Gradient Boosting," is flexible and networked (Kumawat et al., 2021). The best machine learning software for problems like regression, classification, and ranking is available from this package, which also supports parallel tree boosting. It is essential to comprehend the machine learning theories and techniques upon which XGBoost is based, specifically supervised machine learning, decision trees, integrated learning, and gradient boosting.

A model is trained in supervised machine learning using algorithms to find patterns in a collection of labels and features (Sharen & Rajalakshmi, 2022). The labels for the features in a fresh dataset are then predicted using the trained model.

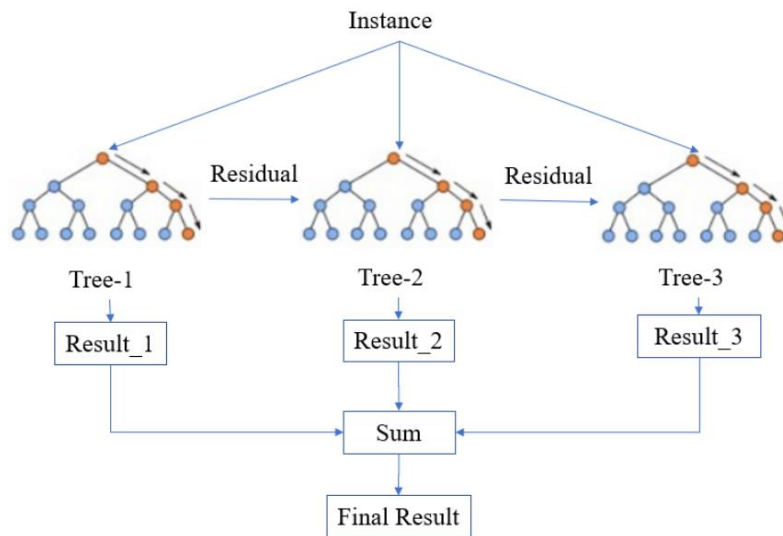


Figure 3.10 Example of XGBoost Classifier

In below figure shows that the formula of XGBoost algorithm. The equation that follows demonstrates how it replicates a particular kind of gradient descent. Since it serves as the loss function to minimise, the indicates the direction in which the function

lowers. The loss function was fitted to the change rate. It is expected to accurately reflect the loss's behaviour and correlates to the gradient descent learning rate.

$$F_{x_{t+1}} = F_{x_t} + \epsilon_{x_t} \frac{\partial F}{\partial x}(x_t)$$

To create an effective function that will converge to the function's minimum and depict the full formula as a sequence, the model must be iterated several times. We utilise this function as an error measure to decrease loss and preserve performance over time. The minimum of the function is finally reached by the sequence. When a gradient boosting regressor is evaluated, the error function is specified using this specific notation.

$$f(x, \theta) = \sum l(F((X_i, \theta), y_i))$$

Figure 3.11 Formula of XGBoost

3.6.1.3 Random Forest

The Random Forest classifier creates several decision trees using a subset of the training dataset that is randomly chosen. Combining the votes from several decision trees yields the final class of test items (Choi et al., 2018). In, a reduced-number-of-trees random forest classification was presented (Priya et al., 2020). In other words, a random forest is a classifier composed of several decision trees trained on various subsets of a dataset and averaged to improve the estimation accuracy of the dataset (Azam et al., 2021). With random forests, the final result is predicted based on the expected majority rather than relying on a single decision tree and estimates from each tree are collected.

A popular ensemble technique that makes advantage of bootstrap sampling is bagging (Access et al., 2021). Bootstrap sampling is used to create several decision trees

using the bagging technique, and the characteristics from the initial sample set are randomly chosen to be utilised for partitioning at each node. Any tree-based algorithm's basic unit is the node. Random Forest reduces the possibility of over fitting brought on by the usage of a single decision tree model. Additionally, the use of bootstrap sampling contributes to the creation of a classification model with improved accuracy and the best possible generalisation capacity (H. Alsagri & Ykhlef, 2020).

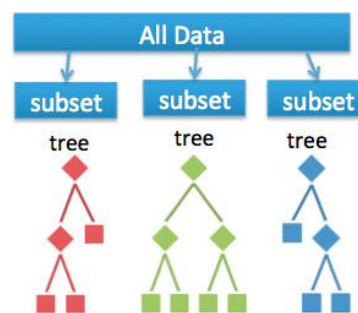


Figure 3.12 Example of Random Forest

3.6.2 Python Machine Learning Libraries

Therefore, to accomplish machine learning tasks using the Python programming language, we need to have access to the Python machine learning libraries. One of the most often used computer languages for this purpose is Python, which has displaced many other languages in the field thanks in part to its enormous library base. Machine learning uses the following Python libraries:

3.6.2.1 Numpy

An essential basic tool for understanding machine learning is the NumPy library. For conducting any mathematical or scientific calculation, several of its capabilities are

highly helpful. Since mathematics is the foundation for machine learning, most mathematical tasks may be accomplished with NumPy.

```
import numpy as np
```

Figure 3.13 Example of Numpy Libraries in command

3.6.2.2 Pandas

Pandas is a well-liked Python module for data analysis. It does not have a direct relationship to machine learning. As we know that the dataset must be prepared before training. In this scenario, Pandas comes helpful as it was created expressly for data extraction and processing. It provides a wide variety of high-level data structures and data analysis capabilities. It comes with a number of integrated methods for collecting, integrating, and filtering data.

```
import pandas as pd # for data manipulation
```

Figure 3.14 Example of Pandas Libraries in command

3.6.2.3 Natural Language Toolkit (NLTK)

In Python, NLTK is a crucial package that enables operations including tokenization, stemming, tagging, parsing, and classification. It essentially serves as our primary tool for machine learning and natural language processing. For Python developers who are just starting out in this profession, it now acts as a foundational teaching resource and machine learning.

```
import nltk # for text manipulation
```

Figure 3.15 Example of NLTK Libraries in command

3.6.2.4 Matplotlib

For data visualisation, Matplotlib is a popular Python library. It is not directly connected to machine learning, unlike Pandas. Particularly when a coder wants to observe the patterns in the data, it is beneficial. To create 2D graphs and charts, a library for 2D charting is used. Plotting data is made easier for programmers by the pyplot module, which provides tools to adjust line styles, font properties, formatting axes, etc. It features several graphs and plots, such as histograms, error diagrams, bar diagrams, and others, for visualising data.

```
import matplotlib.pyplot as plt
```

Figure 3.16 Example of Matplotlib Libraries in command

3.7 SOFTWARE EQUIPMENT

Documentation and development of this project will require the use of multiple software. Software equipment that will be used in this project are shown on the table below:

Software and its specification	Description
Microsoft Office Word 2016	Used to do for report writing documents
Microsoft Office PowerPoint 2016	Used to do the presentation slide
Microsoft Office Excel 2016	Used to store the dataset with format .csv file.

Google Chrome Version 102.0.5005.63	Used to search the research for the project
Jupyter Notebook	Used to allow writing and run the python code.
Python 3.9.2	Used to execute the dataset and run the python code for project

Table 3.1 Software equipment and its specification

3.8 HARDWARE EQUIPMENT

For hardware equipment, it is needed to use highly effective and compatible, as well as have the enough size to download the software, application and documents to prevent problem occur.

Hardware	Specification	Description
Laptop	ASUS VivoBook 15 CPU: Intel(R) Core(TM) i5-8265U CPU @ 1.60GHz 1.80 GHz, 4 Core(s), 8 Logical Processors VGA: NVIDIA GeForce MX250 VRAM: 128 MB	To perform all the documentation and development.
Smartphone	OPPO F9 CPU: Octa Core GPU: Mali-G72 MP3 OS: Android Version 10	Used to find the research information

Table 3.2 Hardware equipment and its specification

3.9 POTENTIAL USE OF PROPOSED SOLUTION

In this research, we can detect that depression tendencies among teenagers are tends to become very serious in real life and make prevention and solutions before it becomes serious. Estimation of the depressive tendencies will conduct by using python language with three machine learning algorithms which are Support Vector Machine, XGBoost, and Random Forest are applied. Dataset will also test on this three machine learning algorithms to compare the accuracy and performance of each algorithm. Through this implementation, we can determine the accuracy result about estimating depressive tendencies of twitter user.

For future work of the research, we may planning try to conduct the same machine learning algorithm to implement the experiment of the estimating depressive tendencies but using different types of datasets to test the accuracy and precision of the result about depressive tendencies. We also intend to look at the efficacy of other characteristics in estimating depression tendencies among Twitter users.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 INTRODUCTION

This chapter focus on discussing about the explanation of the result or outcomes. This chapter will also include the developments that are being implemented utilising resources, techniques, approaches that are appropriate for the research. Last but not least, the component on result and discussion displays the findings that purport to meet the research's goals.

4.2 RESULT

This research is mainly focused on estimating depressive tendencies of twitter user via social media data. Data was downloaded from the Kaggle website that provided the twitter dataset which contains depressive which is 1 and non-depressive tweets which is 0. This twitter dataset was used to estimate the depressive tendencies of twitter user using the Jupyter Notebook with python language and different machine learning model to estimate the result. The result was attained from the development of the machine learning model. Data pre-processing and cleaning were needed to implement in order to convert it become clean of twitter dataset. The sentiment analysis of dataset that in Kaggle website was needed to do some data pre-processing which are tokenization, remove stop words, lower case conversion, stemming, and normalization to become clean dataset. After that, the clean tweet text will visualize the tweets via WordCloud. In WordCloud, the words that are used most often displayed in larger font sizes than the terms that are used less frequently. By using the WordCloud, the most common words in dataset for depressive and non-depressive tweets will be shown. Moreover, dataset will separate into

70% of training and 30% of testing to train and test the Twitter dataset using the different machine learning classifiers which are SVM-Linear, XGBoost, and Random Forest. Accuracy, precision, recall, and f1-score are the four evaluation metrics that will be used to calculate the results to evaluate the performance of the machine learning model. Furthermore, the final outcome of metrics will be used to compare which is the most accuracy with different three machine learning algorithms. The result will show in bar charts and diagrams. Table below show that the outcome with different metrics by using different machine learning classifiers. After that, the most accurate in training and testing will be chosen for implementing the confusion matrix.

Model	Train				Test			
	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
XGBoost Classifier	0.9982	1.0000	0.9920	0.9960	0.9958	0.9985	0.9827	0.9906
Random Forest Classifier	0.9994	1.0000	0.9975	0.9988	0.9961	0.9942	0.9885	0.9913
SVM-Linear Classifier	0.9986	1.0000	0.9938	0.9969	0.9922	0.9799	0.9856	0.9828

Table 4.1 Accuracy, Precision, Recall, and F1-Score Measure on Twitter Depression Dataset Using Different ML Classifiers on Training and Testing

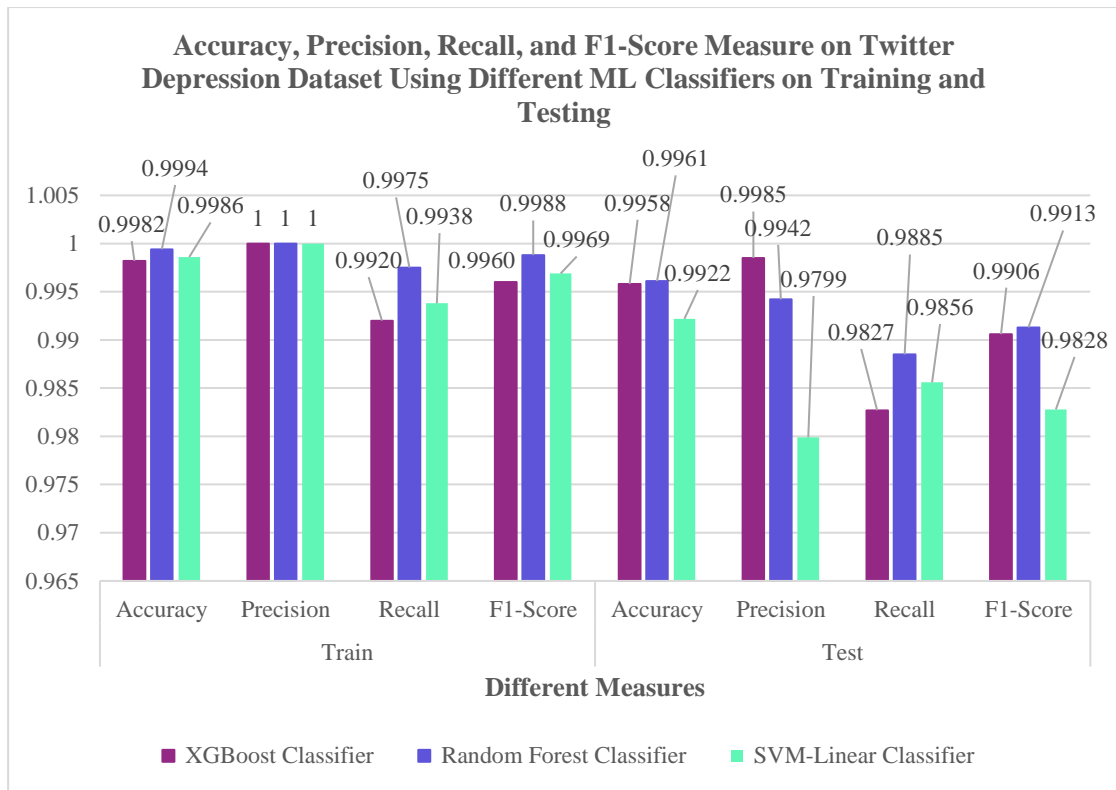


Figure 4.1 Accuracy, Precision, Recall, and F1-Score Measure on Twitter Depression Dataset Using Different ML Classifiers on Training and Testing

In below figure show that the accuracy measure for different classifiers in both training and testing. For training result, the accuracy measure for XGBoost, Random Forest, and SVM-Linear which are 0.9982 (99.82%), 0.9994 (99.94%), and 0.9986 (99.86%). For testing result, the accuracy for XGBoost is 0.9958 (99.58%), Random Forest which is 0.9961 (99.61%), and SVM-Linear which is 0.9922 (99.22%).

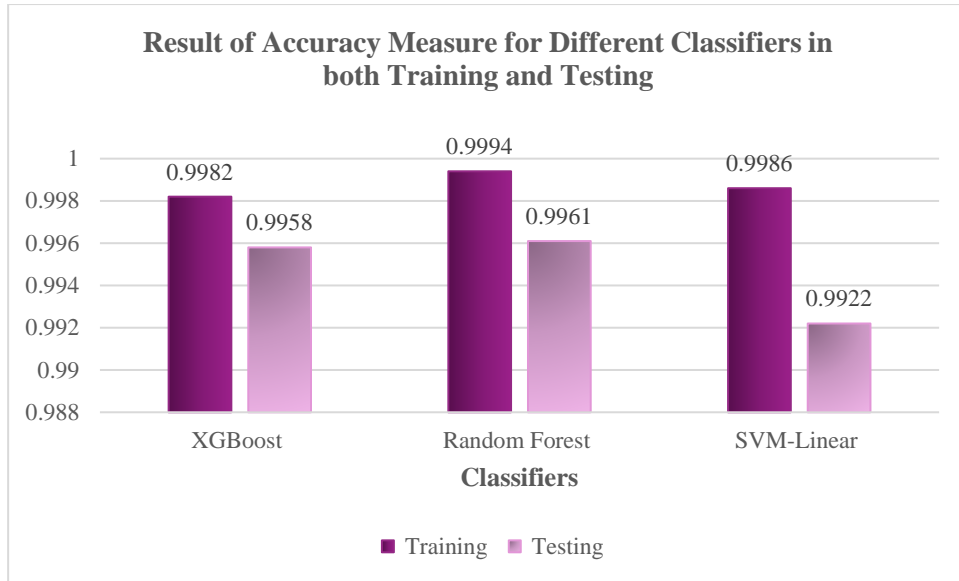


Figure 4.2 Result of Accuracy Measure for Different Classifiers in both Training and Testing

Below figure show that the precision measure for different classifiers in both training and testing. For training result, the precision measure for XGBoost, Random Forest, and SVM-Linear which are both 1.0 (100%). For testing result, the precision for XGBoost is 0.9985 (99.85%), Random Forest which is 0.9942 (99.42%), and SVM-Linear which is 0.9799 (97.99%).

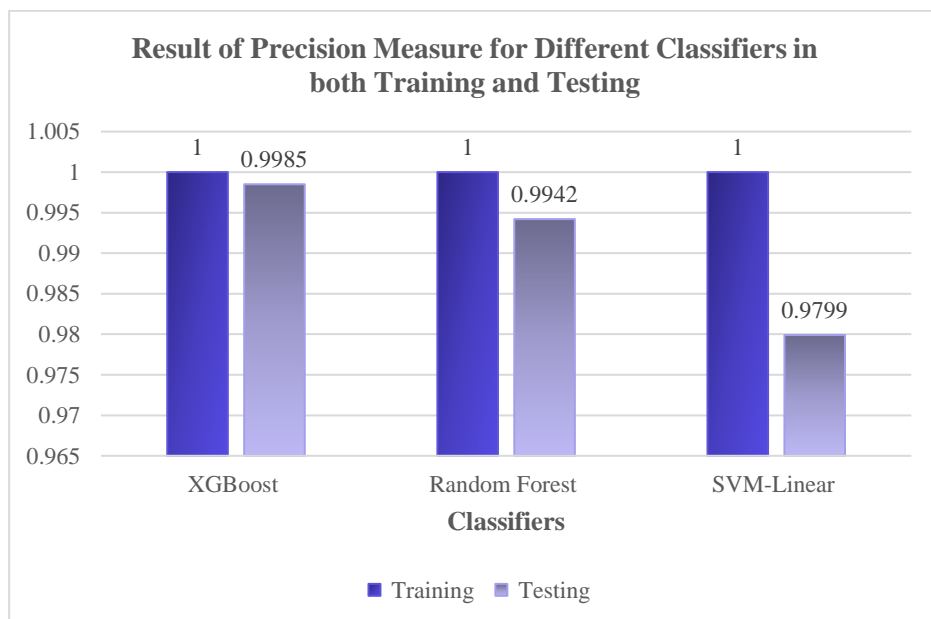


Figure 4.3 Result of Precision Measure for Different Classifiers in both Training and Testing

Below figure show that the recall measure for different classifiers in both training and testing. For training result, the recall measure for XGBoost, Random Forest, and SVM-Linear which are 0.9920 (99.20%), 0.9975 (99.75%), and 0.9938 (99.38%). For testing result, the recall for XGBoost is 0.9827 (98.27%), Random Forest which is 0.9885 (98.85%), and SVM-Linear which is 0.9856 (98.56%).

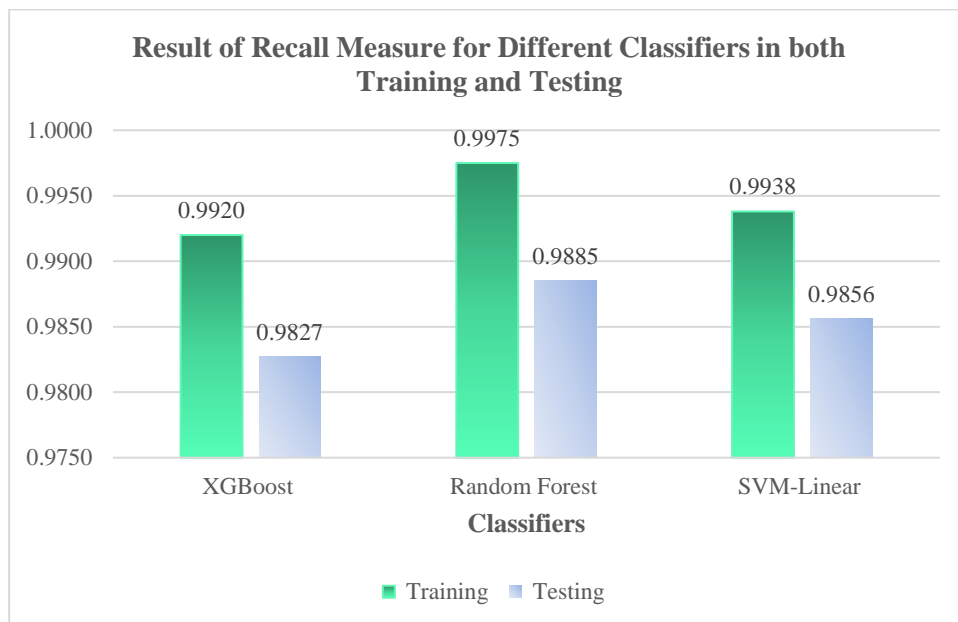


Figure 4.4 Result of Recall Measure for Different Classifiers in both Training and Testing

Below figure show that the f1-score measure for different classifiers in both training and testing. For training result, the f1-score measure for XGBoost, Random Forest, and SVM-Linear which are 0.9960 (99.60%), 0.9988 (99.88%), and 0.9969 (99.69%). For testing result, the recall for XGBoost is 0.9906 (99.06%), Random Forest which is 0.9913 (99.13%), and SVM-Linear which is 0.9828 (98.28%).

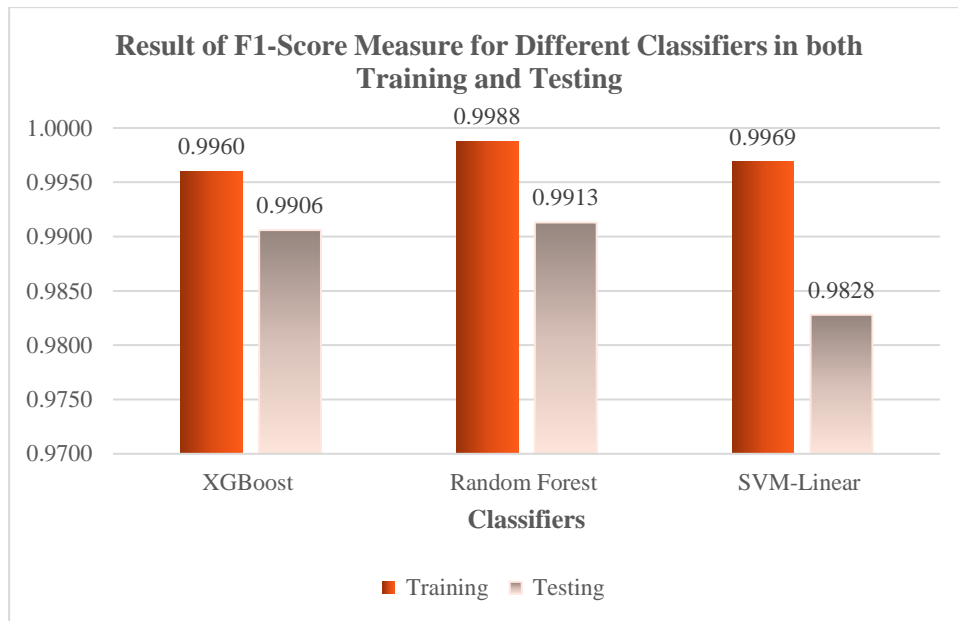


Figure 4.5 Result of F1-Score Measure for Different Classifiers in both Training and Testing

When utilizing word cloud libraries, it has surely noticed a cloud full of variously sized words that represent the frequency or importance of each term. This is known as the Tag Cloud or Word Cloud. The size of each word in a word cloud, a method for visualizing text data, corresponds to its frequency or significance. Use a word cloud to highlight significant textual information. Word clouds are often used in the analysis of data from social networking websites (M. R. Kumar et al., 2022). In Word Cloud, the words that are used most often displayed in larger font sizes than the terms that are used less frequently. By using the Word Cloud, the most common words in dataset for depressive and non-depressive tweets will be shown.



Figure 4.6 All common words in the entire dataset



Figure 4.7 Positive words which is non-depressive tweets in the dataset


```
xgbc = XGBClassifier(max_depth=6, n_estimators=1000, nthread = 3)
```

Figure 4.9 Hyperparameters in XGBoost machine learning classifier

In this project, I employ the hyperparameters max depth, n estimators, and nthread in the XGBoost machine learning method. For max depth, the deepest point a tree can go. When this value is increased, the model becomes more sophisticated and is more prone to overfit. 0 means there is no depth restriction. The max depth's default value which is 6 and the number of trees to be included in the model, n estimators. For nthread, the quantity of simultaneous XGBoost execution threads. In this project, I set the maximum depth at 6, the maximum number of trees to be utilized in the model at 1,000, and the maximum number of concurrent XGBoost threads at 3.

```
rf = RandomForestClassifier(max_depth=None, n_estimators=1000, random_state=42)
```

Figure 4.10 Hyperparameters in Random Forest machine learning classifier

In this project, I also employ the max depth, n estimators, and random state parameters of the Random Forest machine learning approach. The default option for max depth, which determines a tree's maximum depth, is None, which indicates that each tree will grow until every leaf is pure. the number of trees to be included in the model, n estimators. Since 42 is a randomised state, all executions will use the same training and testing sets. To guarantee the reproducibility of the splits created for the random state. The maximum depth in this project is None, the maximum number of trees that may be used in the model is 1000, and the random state is 42.

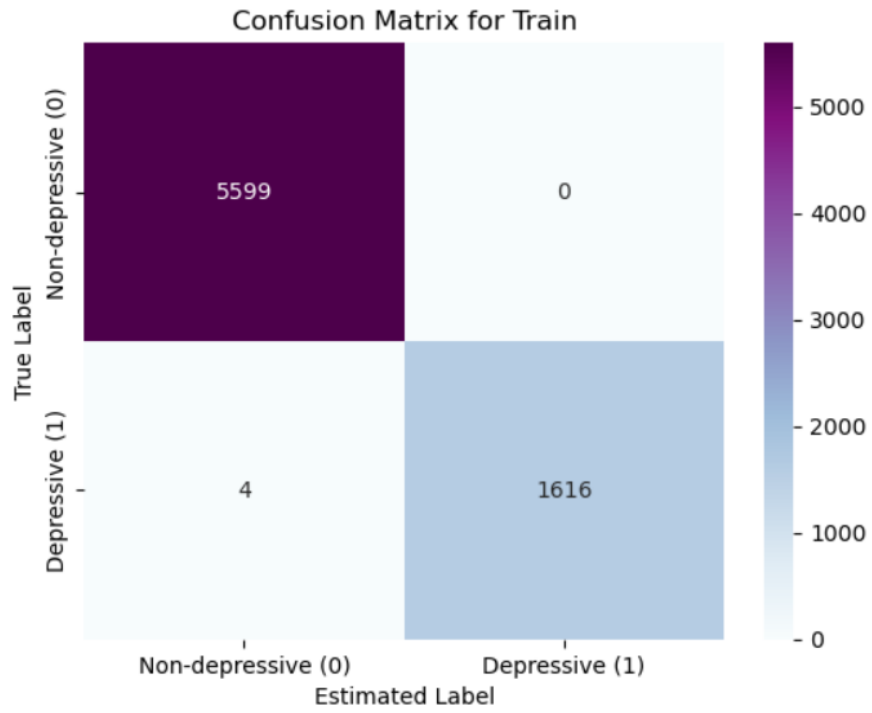


Figure 4.11 Confusion Matrix for Training with Random Forest classifier

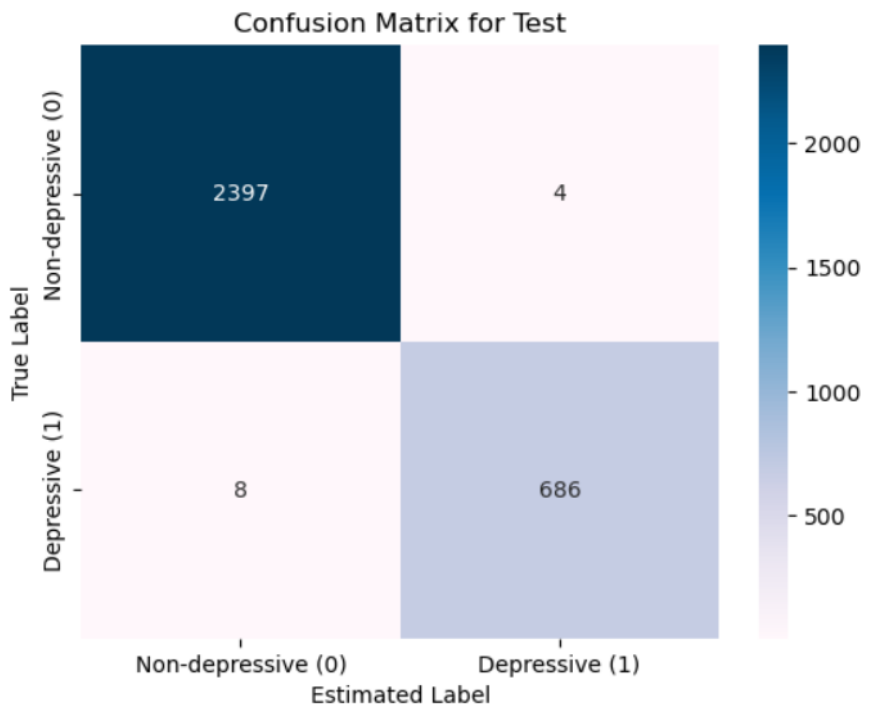


Figure 4.12 Confusion Matrix for Testing with Random Forest classifier

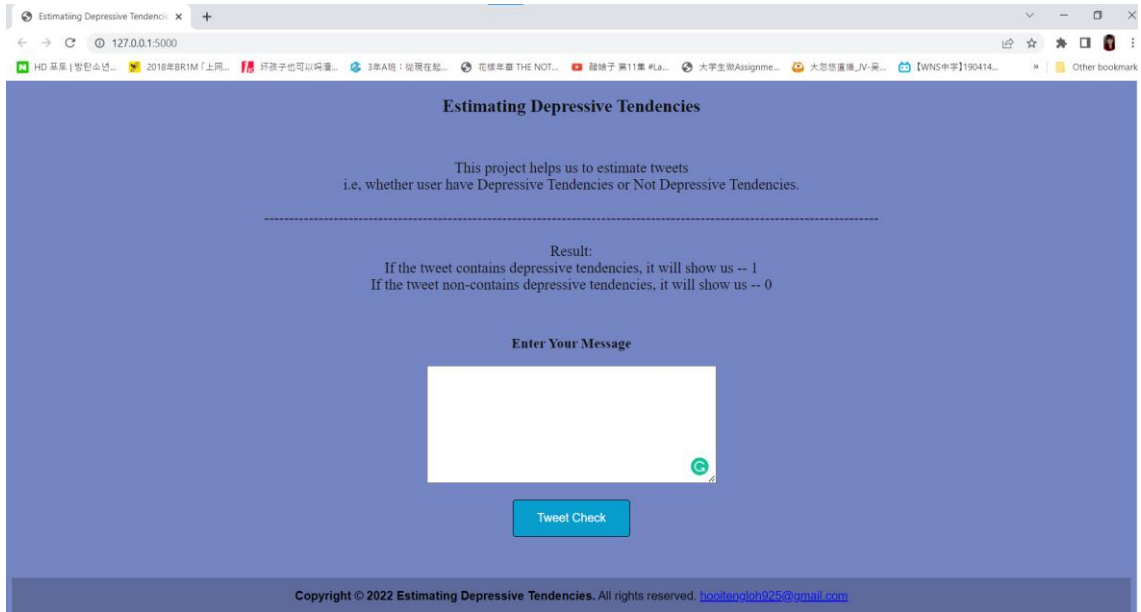


Figure 4.13 Estimating Depressive Tendencies on Tweets

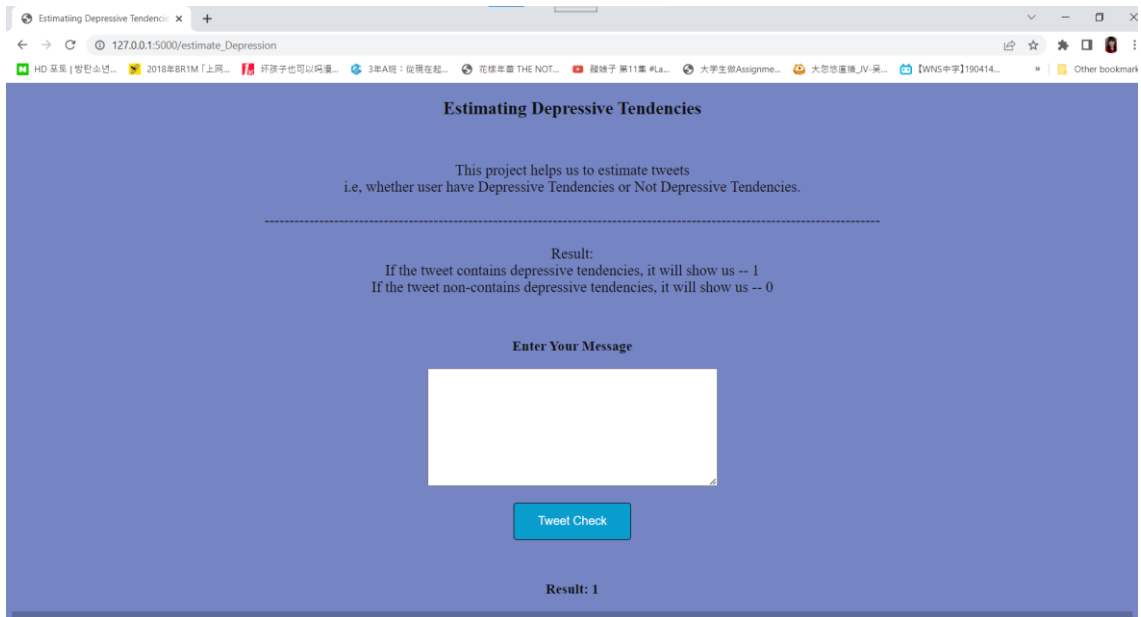


Figure 4.14 Result is 1 (Depressive Tendencies) after inserting the tweet “I am depress”

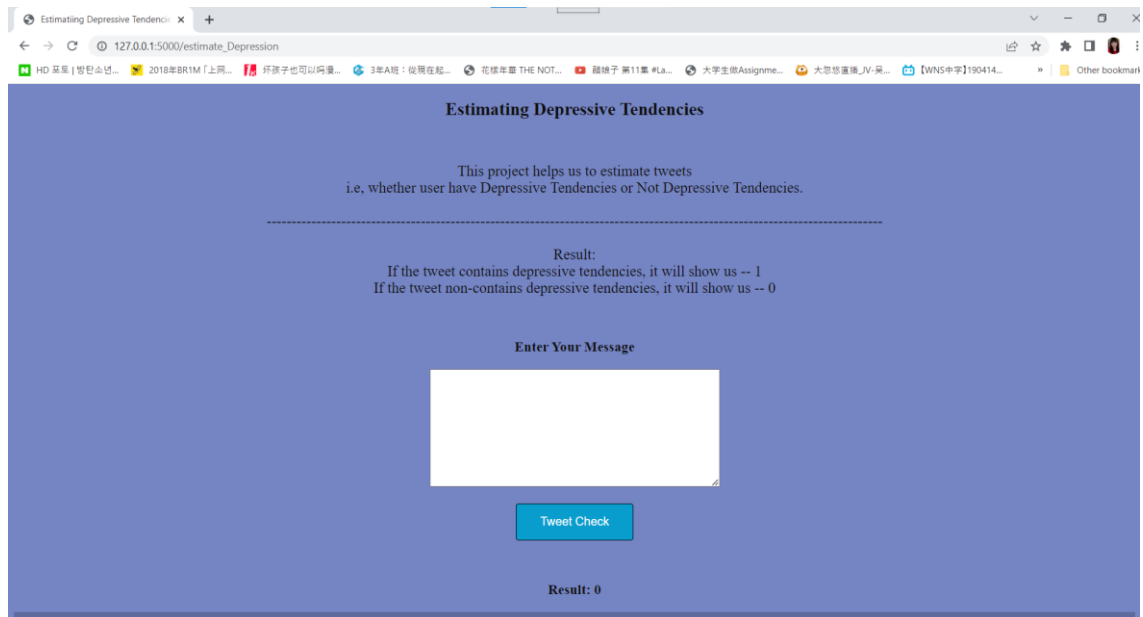


Figure 4.15 Result is 0 (Non-Depressive Tendencies) after inserting the tweet “I am happy”

4.3 DISCUSSION

In conclusion, the best performance for three different machine learning algorithms which is the Random Forest Classifier because the accuracy which is the highest among three machine learning algorithms to estimate the depressive tendencies of twitter user via social media data. Therefore, this research was achieved the objectives which are to study the properties of the text related to depressive tendencies via Twitter dataset, leverage the machine learning algorithm for dataset of the twitter user related to depressive tendencies, and to test the accuracy of the depressive tendencies of twitter user with machine learning techniques for estimating the depressive tendencies.

CHAPTER 5

CONCLUSION

5.1 CONCLUSION

This project which is Estimating Depressive Tendencies of Twitter user via Social Media Data is to detect the depressive tendencies in Twitter dataset. As such, it can estimate the early stages of depressive tendencies before a person is prone to falling into depressive tendencies, causing them to become increasingly inactive thoughts, resulting in suicidal thoughts. Machine learning algorithms were applied to estimate the depression of Twitter user using Twitter dataset. Data were collected from the Kaggle websites that related with the dataset of Twitter user. Subsequently, three machine learning algorithms were applied in this project which are Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), and Random Forest (RF). The Accuracy Score, F1-Score, Precision, Recall, and Confusion Matrix were used to train and test the models. Through the process of data pre-processing, data modelling, data training and testing process, the accuracy of Random Forest (RF) was found to be the highest, although Support Vector Machine (SVM) was identified as the best model. In training process, the accuracy for RF which is 99.94%, for XGBoost and SVM which are 99.81% and 99.86%. In testing process, the best accuracy which is the Random Forest (99.61%), while the XGBoost and SVM which are 99.58% and 99.22%. Therefore, a person's chances of managing depression and minimising its negative consequences on their well-being, health, and social and economic life are improved by early diagnosis and treatment of depressed symptoms. Through machine learning algorithms, we can know if the person is depressed and take action in advance.

5.2 LIMITATION

For limitation is in this research for Twitter dataset is that only 10314 tweets were included in the Twitter dataset that was used. The model's performance will be impacted if there are more than 100,000 tweets. When having the larger Twitter dataset are trained and tested, the result for accuracy will be impacted. Besides that, before the data pre-processing process is implemented, the data is not completely clean yet, it still having some unstructured data and poorly formatted. Therefore, it is needed to do data pre-processing process to ensure that the data are fully clean and might not be affect the result for accuracy.

5.3 FUTURE WORK

Future research can expand on this work by considering more machine learning models that are very unlikely to overfit the data given and by identifying more accurate ways to gauge the influence of the features. Future research can incorporate characteristics for various machine learning classifiers to accurately predict user sadness. Besides that, we can try to implement the machine learning model which can estimate not only depressive or non-depressive but also estimate another level of depression like not depressed, moderately depressed, and severely depressed.

REFERENCES

- [1] Access, O., Joshi, D., & Jain, M. (2021). *STRESS , ANXIETY AND DEPRESSION PREDICTION USING ENSEMBLE*. 12, 357–361.
- [2] Aleem S, Huda N, A. R. et al. (2022). *Machine Learning Algorithms for Depression: Diagnosis, Insights, and Research Directions*.
- [3] Alsagri, H. S., & Ykhlef, M. (2020). Machine learning-based approach for depression detection in twitter using content and activity features. *IEICE Transactions on Information and Systems*, *E103D*(8), 1825–1832. <https://doi.org/10.1587/transinf.2020EDP7023>
- [4] Alsagri, H., & Ykhlef, M. (2020). Quantifying feature importance for detecting depression using random forest. *International Journal of Advanced Computer Science and Applications*, *11*(5), 628–635. <https://doi.org/10.14569/IJACSA.2020.0110577>
- [5] Asad, N. Al, Mahmud Pranto, M. A., Afreen, S., & Islam, M. M. (2019). Depression Detection by Analyzing Social Media Posts of User. *2019 IEEE International Conference on Signal Processing, Information, Communication and Systems, SPICSCON 2019*, 13–17. <https://doi.org/10.1109/SPICSCON48833.2019.9065101>
- [6] Azam, F., Agro, M., Sami, M., Abro, M. H., & Dewani, A. (2021). Identifying Depression among Twitter Users using Sentiment Analysis. *2021 International Conference on Artificial Intelligence, ICAI 2021, Cdc*, 44–49. <https://doi.org/10.1109/ICAI52203.2021.9445271>
- [7] Choi, J., Choi, J., & Jung, H. T. (2018). Applying Machine-Learning Techniques to Build Self-reported Depression Prediction Models. *CIN - Computers Informatics Nursing*, *36*(7), 317–321. <https://doi.org/10.1097/CIN.0000000000000463>
- [8] Hassan, A. U., Hussain, J., Hussain, M., Sadiq, M., & Lee, S. (2017). Sentiment analysis of social networking sites (SNS) data using machine learning approach for the measurement of depression. *International Conference on Information and Communication Technology Convergence: ICT Convergence Technologies Leading the Fourth Industrial Revolution, ICTC 2017, 2017-Decem*, 138–140. <https://doi.org/10.1109/ICTC.2017.8190959>
- [9] Kamde, P. P. M., Ramteke, S., Kadam, R., & Hundiwala, G. (2022). *SVM Classification Technique to Analyze Mental Health and Stress Levels*. 2(7), 412–418. <https://doi.org/10.48175/IJARSCT-4368>
- [10] Kumar, M. R., Pooja, K., Udathu, M., Prasanna, J. L., & Santhosh, C. (2022). Detection of Depression Using Machine Learning Algorithms. *International Journal of Online and Biomedical Engineering*, *18*(4), 155–163. <https://doi.org/10.3991/ijoe.v18i04.29051>
- [11] Kumar, P., Samanta, P., Dutta, S., Chatterjee, M., & Sarkar, D. (2022). Feature Based Depression Detection from Twitter Data Using Machine Learning Techniques. *Journal of Scientific Research*, *66*(02), 220–228. <https://doi.org/10.37398/jsr.2022.660229>
- [12] Kumawat, K., Kumawat, G., Chakrabarti, P., Poddar, S., Chakrabarti, T., Hussaine, J., Kamali, A., Bolsev, V., & Kateb, B. (2021). A Machine Learning Technique to

Analyze Depressive Disorders. *Research Square*, 1–10.

- [13] Li, W., Wang, Q., Liu, X., & Yu, Y. (2021). Simple action for depression detection: using kinect-recorded human kinematic skeletal data. *BMC Psychiatry*, 21(1), 1–11. <https://doi.org/10.1186/s12888-021-03184-4>
- [14] Nadeem, M. (2016). *Identifying Depression on Twitter*. 1–9. <http://arxiv.org/abs/1607.07384>
- [15] Neelavathi, G., Sowmiya, D., Sharmila, C., & Vaishnavi, J. (2021). Sentiment Analysis for Depression Based on Social Media Post by Using Natural Language Processing. *International Journal of Advanced Research in Science, Communication and Technology*, 12(2), 134–139. <https://doi.org/10.48175/ijarsct-2319>
- [16] Priya, A., Garg, S., & Tigga, N. P. (2020). Predicting Anxiety, Depression and Stress in Modern Life using Machine Learning Algorithms. *Procedia Computer Science*, 167(2019), 1258–1267. <https://doi.org/10.1016/j.procs.2020.03.442>
- [17] Reece, A. G., Reagan, A. J., Lix, K. L. M., Dodds, P. S., Danforth, C. M., & Langer, E. J. (2017). Forecasting the onset and course of mental illness with Twitter data. *Scientific Reports*, 7(1), 1–11. <https://doi.org/10.1038/s41598-017-12961-9>
- [18] Sawangarreerak, S., & Thanathamthee, P. (2020). Random forest with sampling techniques for handling imbalanced prediction of university student depression. *Information (Switzerland)*, 11(11), 1–13. <https://doi.org/10.3390/info11110519>
- [19] Sharen, H. G., & Rajalakshmi, R. (2022). DLRG@LT-EDI-ACL2022: Detecting signs of Depression from Social Media using XGBoost Method. *LTEDI 2022 - 2nd Workshop on Language Technology for Equality, Diversity and Inclusion, Proceedings of the Workshop*, 346–349. <https://doi.org/10.18653/v1/2022.ltedi-1.53>
- [20] Singh, S., & Choudhary, S. S. (2017). Social Media Data Analysis: Twitter Sentimental Analysis Using R Language. *International Journal of Advances in Electronics and Computer Science*, 4(11), 13–17. <http://iraj.in>
- [21] Tsugawa, S., Mogi, Y., Kikuchi, Y., Kishino, F., Fujita, K., Itoh, Y., & Ohsaki, H. (2013). On estimating depressive tendencies of Twitter users utilizing their tweet data. *Proceedings - IEEE Virtual Reality*. <https://doi.org/10.1109/VR.2013.6549431>

APPENDIX A

Gantt Chart

