PREDICTING SUICIDAL IDEATION VIA SOCIAL MEDIA

BOON KAR LIH

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

UNIVERSITI MALAYSIA PAHANG

UNIVERSITI MALAYSIA PAHANG

DECLARATION OF THESIS AND COPYRIGHT

Author's Full Name : BOON KAR LIH

Date of Birth :

Title : PREDICTING SUICIDAL IDEATION VIA SOCIAL MEDIA

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)

New IC/Passport Number
Date: 19 January 2023

_____
(Supervisor's Signature)

Dr. Nur Shazwani binti Kamarudin

Name of Supervisor
Date: 19 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: 19 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.

(Supervisor's Signature)

Full Name : Dr. Nur Shazwani binti Kamarudin

Position : Senior Lecturer

Date : 19 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.

karlih

_____

(Student's Signature)

Full Name : BOON KAR LIH

ID Number : CA19046

Date : 19 January 2023

PREDICTING SUICIDAL IDEATION VIA SOCIAL MEDIA

BOON KAR LIH

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

First, 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.

Then, I would like to thanks to all who have been give help to me throughout this final year project. Especially to my research supervisor, Dr. Nur Shazwani binti Kamarudin who has been guiding throughout the process.

Not to forget, I am extending my heartfelt gratitude to my parents and my friends who has been supporting me. The completion of this research project may not be achievable without the support from them.

Finally, I want to express my gratitude to everyone who helped me out with hands, support, and time, both directly and indirectly.

ABSTRAK

Pada masa kini, bunuh diri adalah salah satu punca utama kematian di seluruh dunia, dengan lebih 800,000 orang mati akibat bunuh diri setiap tahun. Fikiran bunuh diri adalah renungan dan keasyikan tentang bunuh diri. Kebanyakan orang yang mempunyai idea untuk membunuh diri aktif dalam media sosial dan menghantar tanda tentang niat mereka. Walau bagaimanapun, pengelas yang tepat boleh mengenal pasti data yang mungkin memberi petunjuk ke arah idea bunuh diri. Matlamat penyelidikan ini adalah untuk mengkaji idea bunuh diri melalui Subreddits pada dataset Reddit. Set data dikumpulkan daripada tapak web Kaggle yang mana set data dikumpulkan dari tahun 2008 hingga 2021. Model ini akan meramalkan jika individu tersebut mempunyai idea untuk membunuh diri atau tidak membunuh diri berdasarkan set data. Tiga algoritma Pembelajaran Mesin dilaksanakan untuk meramalkan hasil dan hasil iaitu Mesin Vektor Sokongan (SVM), Pokok Keputusan dan Naive Bayes (NB). SVM memberikan hasil yang paling tepat untuk ketepatan dengan 93.40% antara yang lain. Ia boleh meramalkan idea bunuh diri dengan tepat melalui data media sosial.

ABSTRACT

Nowadays, suicide is one of the leading causes of mortality worldwide, with over 800,000 people dying by suicide each year. Suicidal ideation is a contemplations and preoccupations about suicide. Most of the people who got suicidal ideation are active in social media and send out signs about their intentions. However, an accurate classifier can identify the data which may potentially hint towards suicidal ideation. The aim of this research is to study the suicidal ideation via Subreddits on Reddit dataset. The dataset is collected from Kaggle websites which dataset is collected from 2008 until 2021. The model will predict if the individual has suicidal or non-suicidal ideation based on the dataset. Three Machine Learning algorithms are implemented to predict the result and outcome which are Support Vector Machine (SVM), Decision Tree and Naive Bayes (NB). SVM give the most precise result for accuracy with 93.40% among the others. It can accurately predict the suicidal ideation via social media data.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

SVM	Support Vector Machine
NB	Naïve Bayes

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

Social media has evolved into a part of our everyday life. Majority people nowadays are active on popular social media platforms such as Facebook, Instagram, Twitter, Reddit and others. They post about their daily life such as their food, their trips, and their outfit of the day. Reddit is a prominent online community for frank talks on taboo topics like suicidal ideation. The posts are organised into communities or subreddits, which are user-created boards, based on their subject. Social networks, on the other hand, can reflect a variety of social phenomena such as depression, suicide, and so on. People will also use social media to express their feelings. Addiction to drugs and alcohol, anxiety, sadness, hopelessness, and other conditions are also risk factors for suicide (Ribeiro et al., 2012). Most people who got suicidal ideation are active in social media and send out signs about their intentions. For example, they will post some statements like “I hate my life” which are a precursor to suicide (Mbarek et al., 2019). A study analysed tweets from a young girl who had just committed suicide and had posted them 24 hours earlier. (Poulin et al., 2014).

According to the World Health Organization (WHO), it is estimated that 800,000 people die by suicide each year worldwide with at least as many suicide attempts (Vioules et al., 2018). This drives WHO guidelines to assist the globe in meeting the goal of halving the suicide rate by 1/3 by 2030. Suicidal ideation, often known as suicidal thoughts or ideas which refers to a variety of preoccupations about death and suicide. There are two categories of Suicidal ideation which are passive and active. Passive ideation happens when a person wishes to die or considers suicide but does not intend to do so. Suicidal ideation that is active includes not only thinking about suicide but also planning to do so and making a strategy to do so.

However, suicide can be avoided by analysing such posts on social networking platforms (Patel & Soni, 2021). Using social media has opened new opportunities for scholars to use automated language analysis techniques to improve comprehend the thoughts, feelings, beliefs, conduct, and personalities of people (Braithwaite et al., 2016). Reddit data is important in studying suicidal thoughts since it is a gold mine of data because practically all user posts are public and pullable. Reddit data set can be obtained by the Kaggle website. In a another study, Reddit is used to examine how the language used in user comments on the discussion board affects people's propensity for suicidal ideation (De Choudhury & Kiciman, 2017).

Natural language processing (NLP) combines computational linguistics with statistical, machine learning, and deep learning models. These technologies enable computers to comprehend and process human language in the form of text or speech data, including the purpose and sentiment of the speaker or writer (Hassan et al., 2017). The python programming language will be used in this study, as it offers a large selection of tools and libraries for tackling specific NLP tasks. The Natural Language Toolkit (NLTK) is an open-source collection of libraries, tools, and training resources for the development of natural language processing (NLP) programmes, contains several of them. Dealing with a large amount of data set, sentiment analysis will be applied and solely focus on suicidal ideation in Subreddit of Reddit.

1.2 PROBLEM STATEMENTS

Nowadays, billions of people like to use social media to express their feelings no matter good or bad. However, when the people with suicidal ideation write some posts of emotionally instead of getting real help from therapist, it is harmful for them. They cannot get any help from social networks as humans require face-to-face contact. People who place a higher value on social media than in-person relationships are more likely to acquire or aggravate mood disorders including anxiety and melancholy.

Other than that, if people write posts with sign of suicidal ideation and seek for help, there will be no professional person such as therapist answer them. Some netizens' joking respond may make that people more depress and more deeply suicidal ideation as they do not get any positive thinking and supports from netizens. Besides, when these group of people saw other people negatively posts, this will make their suicidal ideation become more deeply because they gain more negative mood. There may be some online support groups, but they are not the same as therapy. The best way is to see a therapist and get helpful advice and feedback.

1.3 OBJECTIVE

This study's goal is to predict the suicidal ideation via Subreddits on Reddit. Hereby, three objectives have been identified.

I. To study and identified features based on social media dataset.
II. To analyse available text analysis methods for predicting suicidal ideation.
III. To implement a machine learning algorithm based on the features that has been identified.

1.4 SCOPE

The scopes are listed as follows to get the study results:

Research scope:

I. The study used a dataset from a social media platform, Kaggle (Subreddit of Reddit).
I. Datasets are split into training (70%) and testing (30%).
II. The testing dataset will be run three times.
III. The result will show the accuracy of the sentiment suicidal ideation posts.

1.5 REPORT ORGANIZATION

This thesis is composed of three chapters: introduction, literature review, methodology. Chapter 1 discusses about the introduction of the research. It introduces the research project, problem statement, objectives, scope and significance of the project.

Chapter 2 discusses the literature review of the research. It discusses and compares the previous research about the study, technique used, advantages and disadvantages of existing research.

Chapter 3 explains the methodology of the proposed research. This chapter explain the research method for the study and potential proposed solution in real-time solution.

Chapter 4 discusses the results and discussion of the project. It discusses the result obtained from the implemented method and technique.

Chapter 5 is focus on the conclusion part. It explains the limitations and future work for this research.

CHAPTER 2

LITERATURE REVIEW

2.1 INTRODUCTION

Suicidal ideation research is a big focus, even more so than that of the internet. Predicting suicidal ideation via social media has become an increasingly important research area. This section discusses the relationship between social media and suicide ideation, as well as suicidal prediction methodologies and associated research, as well as the strategies employed.

2.2 PRIOR RESEARCH WORK

This study (Sakib et al., 2021) propose an analysis of suicidal tweets from Twitter using ensemble machine learning methods. Its main goal is to improve the current suicide ideation prediction algorithm's effectiveness and develop a new one. The data set of this study is collected from Twitter through a public API named “tweepy”. The dataset is consisting of text which contains negative word such as “kill myself”, “end my life” and others. The 80% of the data for training and the 20% for testing. Then, the dataset is labelled as 0 and 1 which label as “suicidal” and “non-suicidal”. Before proposed the machine learning algorithm, the dataset go for data pre-processing first as it contains emojis, hashtags and grammatical errors. NLTK is used for pre-processing and data cleaning. Firstly, to remove the hashtags, then removed all the digits and convert to all lower-case texts. Next step undergo tokenization and lemmatization by using NLTK library. Lastly, stop words are removed from dataset also by using NLTK library. Support Vector Machine (SVM), Decision Trees, Logistic Regression, Naïve Bayes, K-Nearest Neighbors, AdaBoost, Catboost, XGBoost, Gradient Boost, and Bagging classifier are used to analyse the suicidal ideation related tweets. Parameter used were default to initialize the estimators. They discovered that the Logistics Regression, Support Vector Machine and Decision Tree Classifier gives the best results among others. Logistics Regression give 95.3% accuracy, 95.8% precision, 94.8% recall and 95.2% F1_score while Support Vector Machine (SVM) give 86.3% accuracy, precision of 96.7%, recall

of 95.9% and F1_score of 96.2%. Decision Tree give 88.5% accuracy, 90.3% precision, 87.2% recall and 88.0% F1_score. Then, they implement voting classifier from sklearn to combine the three machine learning models. The limitation for this research is the dataset can be more professional as it gives more accuracy. Besides, the twitter dataset used is restricted to only 180 characters. It may make the performance of the model become less.

In (S. Jain et al., 2019) the authors propose a machine learning based depression analysis and suicidal ideation detection system. The data set in this study is collect from two ways which are questionnaire and Twitter. Questionnaire dataset is contained 18 features. They're obtaining Twitter data from Reddit, which stands for Python Reddit API wrapper. The data set split to 80-20 where 80% is for training and 20% is for testing. After that, dataset has been removed excess white space and convert into lower case. This study also using NLTK for removal all non-alphabetic characters and stop words. Besides, stemming also conducted for the feature reduction. Feature extraction for dataset I is using LabelEncoder to transform categorical labels to numerical labels. While for dataset II, Tf-idf (Term Frequency Inverse Document Frequency) is used to weight word count feature extraction to for feature vectors. Then, supervised algorithms which include Logistic Regression, Decision tree classifier and XGBoost algorithm are used to determine suicidal ideation with dataset. Test data sets which already undergo pre-processing is used to test the classifiers. The performance of the machine learning classifier is in terms of accuracy, precision, recall and F-measure. Result show that Logistic Regression Classifier is the most accurate of 86.45%. The limitation in this research is when the feature of the dataset is more, the Logistic Regression classifier will starts to falter. Therefore, the features and categorical variables of dataset should be reduced to obtain a more accuracy prediction result.

Another work by (Patel & Soni, 2021) studied the machine learning based approach for prediction of suicide related activity. Machine learning-based approaches such as Support Vector Machine (SVM), Naïve bayes (NB), and Random Forest (RF) used in this study. This research is used Twitter dataset which is collected from Kaggle. The dataset is consisting of text, emoji and hashtags. Then, tokenization and feature extraction are applied for data cleaning. The tweets categorized to 0 (negative), 2 (neutral) and 4 (positive). The N-gram model is then used to find the score using a dictionary.

Extra tree classification is used along with Support Vector Machine (SVM), Naïve bayes (NB), and Random Forest (RF) in this study. Extra Trees is a machine learning technique that integrates many decision tree forecasts into a single forecast. The result outcome with accuracy of SVM is 94%, Decision Tree is 89%, Naïve bayes is 95%, Random Forest is 90% and Extra tree is 96%. The study found that Extra tree is the most accuracy of 96% across all classifications.

2.3 COMPARISON TABLE BETWEEN EXISTING RESEARCH WORKS

Elements	Research 1	Research 2	Research 3
Research and Author	Analysis of Suicidal Tweets from Twitter Using Ensemble Machine Learning Methods (Sakib et al., 2021)	A Machine Learning based Depression Analysis and Suicidal Ideation Detection System using Questionnaires and Twitter (S. Jain et al., 2019)	Machine Learning Based Approach for Prediction of Suicide Related Activity (Patel & Soni, 2021)
Domain	Improve the existing suicide prediction algorithm's performance and suggest a new one.	Predict the suicidal acts based on the depression levels.	Identify and categories suicide cases by using Machine Learning Based Approach.
Technique	• Pre-Processing: NLTK • Machine Learning Algorithm: Support Vector Machine, Decision Trees, Logistic Regression, Naïve Bayes, K-Nearest Neighbors, AdaBoost,	• Pre-Processing: NLTK • Feature Extraction: Term Frequency Inverse Document Frequency • Machine Learning Algorithm: Logistic Regression. Random Forest	• Tokenization • Feature Extraction • N-gram Model • VADER-Laxicon Score Dictionary • Machine Learning Algorithm: Support Vector Machine, Decision Tree, Navier buyers,

	Catboost, XGBoost, Gradient Boost and Bagging	Classifier, XGBoost Classifier, Support Vector Machine	Random Forest and Extra tree classifier
Data	Tweepy	Questionnaires and Twitter dataset from Reddit	Twitter dataset from Kaggle
Advantages	Logistic Regression, Support Vector Machine and Decision Tree Classifier gives the best results among all the combination. Then, voting classifier is used to merges the three classifier together for predicting the result.	Logistic Regression is the most accurate in detect suicidal ideation. Logistic regression is more straightforward to apply, analyse, and train. It has a high level of accuracy for a variety of simple data sets.	Extra Trees give the highest accuracy. In terms of performance, Extra Trees are on par with or better than the random forest algorithm. It's easy to set up, requiring only a few critical hyperparameters and sound heuristics. It is more efficient and saves time.
Disadvantages	The model interpretability decreases, as one cannot interpret the model using shap, or lime packages.	Logistic regression starts to fail when there are a lot of features and a lot of missing data.	Extra Tree Classifier randomness doesn’t come from bootstrap aggregating but comes from the

			random splits of the data.
Limitation	Twitter dataset restricted to only 180 characters. Large and unprofessional dataset will make the accuracy fall.	Large number of features and too many categorical variables will make machine learning algorithm start to falter.	Large number of features which will affect the model accuracy

Table 2.1 Compare between prior studies

2.4 SUMMARY

In summary, different techniques used in previous research have its own pros and cons. The most reliable technique is Support Vector Machine and Decision Tree Classifier as it comes with the good performance of accuracy, precision, recall and F1_score result compared to others. Besides, it's advantage is more than disadvantage.

CHAPTER 3

METHODOLOGY

3.1 INTRODUCTION OF RESEARCH FRAMEWORK

This chapter explain the method and technique use to predict the suicidal ideation via social media, Reddit. Machine learning technique will be use in this project to fulfil the objective as mentioned before. The machine learning algorithm include Support Vector Machine (SVM), Decision Tree and Naïve Bayes (NB). Support Vector Machine (SVM) is chosen as it can handle both classification and regression on linear and non-linear data. It finds more accurate results because of their ability to handle complex data. For naïve bayes, it is simple and easy to implement as it is fast and can be used to make real-time prediction. Besides, it is not sensitive to irrelevant features. Last but not least, main benefits of using a decision tree is its simplicity as the decision-making process is easy to visualize and understand.

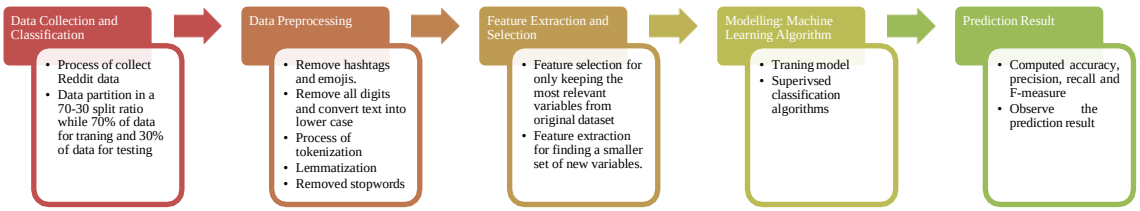


Figure 3.1 Research Framework

3.2 DATA COLLECTION AND CLASSIFICATION

Reddit data will be collected from Kaggle as it is a usable data set which published by previous study. The data collected is a collection of posts from "SuicideWatch" which is a subreddits of the Reddit Platform through Kaggle. "Suicide Watch" is the name of a surveillance process used to deter suicide attempts (Aldhyani et al., 2022). The data set from Kaggle will be used to predict suicidal ideation via social media in this project. It is a text data set with total number of 233440 and divide into three columns. First is number of data, followed by text content and third one is class which divide into suicidal and non- suicidal.

Data set is partitioned in a 70-30 split where 70% of the data is for training and 30% is for testing. Therefore, there will be 163408 data for training and 70032 data for testing.

3.3 DATA PREPROCESSING

Data collected may not be structured such as contain special symbol, hashtags, emojis, grammatical errors, and stop word. All these need to be cleaned and get it into a structed format. Natural Language toolkit (NLTK) stop words corpus is utilised for removal stop word in this research. Tokenization will be accomplished by breaking down a stream of text into separate tokens using the NLTK library. Normalization will reduce word inflections into a lemma, a generic root (Yeskuatov et al., 2022). Tokenization and normalization used to remove all non-English, lower case all posts, convert hash tags to regular words, eliminating any unnecessary letters, spaces and so on. Finally, the dataset has been cleaned and is ready to be vectorized.

3.4 FEATURE EXTRACTION AND SELECTION

Feature extraction is a process to reduce the original set of data to more manageable groups for processing without losing important or relevant information. Undergo feature extraction is because if there is a large data sets with large number of variables will require a lot of computing resources to process. As with feature selection which used for reducing the number of features from original features set to reduce model complexity and enhance model computation efficiency. For both feature extraction and selection, it will use python library which is Scikit Learn for processing.

3.5 MACHINE LEARNING ALGORITHM

Machine Learning Algorithm used in this project are Support Vector Machine (SVM), Decision Tree and Naïve Bayes (NB). Support Vector Machines are a type of machine learning that works with data that has been labelled (T. Jain et al., 2021). Decision Tree is a tree like structure classifier which can be used in both classification or regression (T. Jain et al., 2021). The decision tree approach uses a tree data structure to

make judgments at several levels, making it ideal for prediction problems (Priya et al., 2020). Third algorithm, Naïve Bayes (NB) is also a supervised machine learning algorithm which input data type is labelled. Naïve Bayes (NB) is based on Bayes algorithm and works on conditional probability. (T. Jain et al., 2021).

3.6 PREDICTION RESULT

Prediction results are conducted using machine learning algorithm for training and testing data. The result will contain accuracy, precision, recall and F1_score. The purposeful value's accuracy is determined by how close it is to the genuine value. The degree to which two or more measurements are near to each other is referred to as precision. As a consequence of the calculation, it is defined as the proportion of objects that are effectively classified into a specific class. The number of actual positives computed by recall shows the level of classification that was effectively classified. The consonant mean of the two is F1 score, which addresses a balance between the two (Sakib et al., 2021). Results will be in table, charts and graph. The most accurate algorithm will be obtained through comparison between the three algorithm which are Support Vector Machine (SVM), Decision Tree and Naïve Bayes (NB).

Support Vector Machine				
Class	**Precision**	**Recall**	**F1-Score**	**Support**
Positive	0.99	0.91	0.95	150
Neutral	1.00	0.88	0.94	113
Negative	0.86	1.00	0.92	159
Decision Tree				
Class	**Precision**	**Recall**	**F1-Score**	**Support**
Positive	0.77	1.00	0.87	150
Neutral	1.00	0.75	0.86	113
Negative	1.00	0.89	0.94	159
Naïver buyers				
Class	**Precision**	**Recall**	**F1-Score**	**Support**
Positive	0.97	0.97	0.97	150
Neutral	1.00	0.85	0.92	113
Negative	0.90	0.99	0.94	159

Figure 3.2 Sample Prediction Result

3.7 PROJECT REQUIREMENT

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

3.7.1 INPUT

Input will be the Reddit data which collected from Kaggle. The data collected is a collection of posts from “SuicideWatch” which is a subreddits of the Reddit Platform through Kaggle. The data set from Kaggle will be used to predict suicidal ideation via social media in this project. It is a text data set with total number of 233440 and divide into three columns. First is number of data, followed by text content and third one is class which divide into suicidal and non-suicidal.

3.7.2 OUTPUT

Output will show the accuracy, precision, recall and F1_score of the three algorithms about the prediction of suicidal ideation via social media. First parameter is accuracy. Formula of calculate accuracy is:

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions}$$

Figure 3.3 Formula of Accuracy

However, in binary classification case, accuracy is express in True / False / Positive / Negative values.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

		Predicted	
		Negative	**Positive**
Actual	**Negative**	True Negative	False Positive
	Positive	False Negative	True Positive

True Positives : The cases in which we predicted YES and the actual output was also YES.

True Negatives : The cases in which we predicted NO and the actual output was NO.

False Positives : The cases in which we predicted YES and the actual output was NO.

False Negatives : The cases in which we predicted NO and the actual output was YES.

Figure 3.4 Formula of Accuracy in binary classification

Next, precision refers to how precise or accurate the model is in terms of how many of the anticipated positives are actually positive. It is in the proportion of correct predictions in a class out of all forecasts in that class. Formula of Precision:

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$= \frac{True\ Positive}{Total\ Predicted\ Positive}$$

Figure 3.5 Formula of Precision

Third parameter is Recall which by identifying it as Positive, Recall estimates how many of the Actual Positives the model captures (True Positive). Recall will be the model metric used to identify the optimal model when False Negative has a high cost. Formula of Recall:

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$= \frac{True\ Positive}{Total\ Actual\ Positive}$$

Figure 3.6 Formula of Recall

Where Total actual Positive = True Positive + False Negative

Last measurement is F1_score which is used to seek a balance between Precision and Recall. The higher the F1 Score, the better the model's performance. Formula of F1_score:

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

Figure 3.7 Formula of F1_score

The three algorithms are Support Vector Machine (SVM), Decision Tree and Naïve Bayes (NB). Then, the most suitable algorithm will provide the most accurate result for the prediction of suicidal ideation.

Confusion Matrix is used to summarize the performance and prediction results of the most accurate Machine Learning Algorithm. Count values are used to describe the number of accurate and inaccurate predictions for each class. It is a table with 4 different combinations of predicted and actual values.

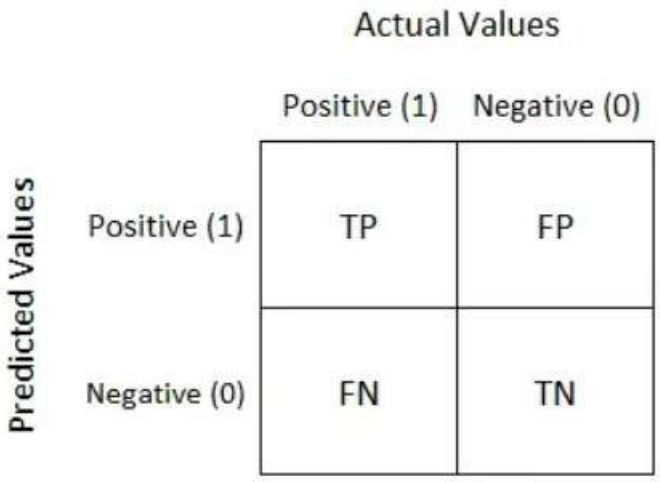


Figure 3.8 Confusion Matrix

3.7.3 PROCESS DESCRIPTION

Process involved in this project include 5 stages which are data collection, data pre-processing, data classification, modelling: Machine Learning Algorithm and predict result. Firstly, collect Reddit data from Kaggle website which provide dataset that relate to suicidal ideation. Next, do data pre-processing. Data cleaning and tokenization will be done in this step. The data which not in a structured form will be cleaned and get it into a structured format. Natural Language toolkit (NLTK) will be used to remove the stop word while also used for tokenization and normalization. Tokenization and Normalization will remove any non-English characters, lowercase all posts, convert hash tags to regular words, and eliminate any unnecessary letters, spaces, and so on. After the data cleaned, the dataset can be used for classification. Data set will be partitioned in a 70-30 split where 70% for training and 30% for testing. Machine learning algorithm is used to analyse the data and get the most accurate result within three different algorithm which are Support Vector Machine (SVM), Decision Tree and Naïve Bayes (NB). The result with highest accuracy will be the most accurate algorithm.

3.7.4 CONSTRAINTS AND LIMITATIONS

The constraint in this research is not involve all the machine learning algorithm. Machine learning algorithm are classified into four types which are supervised, unsupervised, semi-supervised and reinforcement. Each classes contain various type of algorithm such as Linear regression, logistic regression, decision tree, KNN algorithm and K-means. Therefore, it become a constraint to identify which algorithm is the best and accurate for this research.

The limitation in this research is the machine learning technique may not so accurate when the dataset is too big. Besides, one of the limitations is the reddit dataset restricted to only 180 characters. Large number of features and too many categorical variables will also make the time consume become longer and the result may have some small inaccurate.

3.8 PROPOSED DESIGN

3.8.1 Flowchart

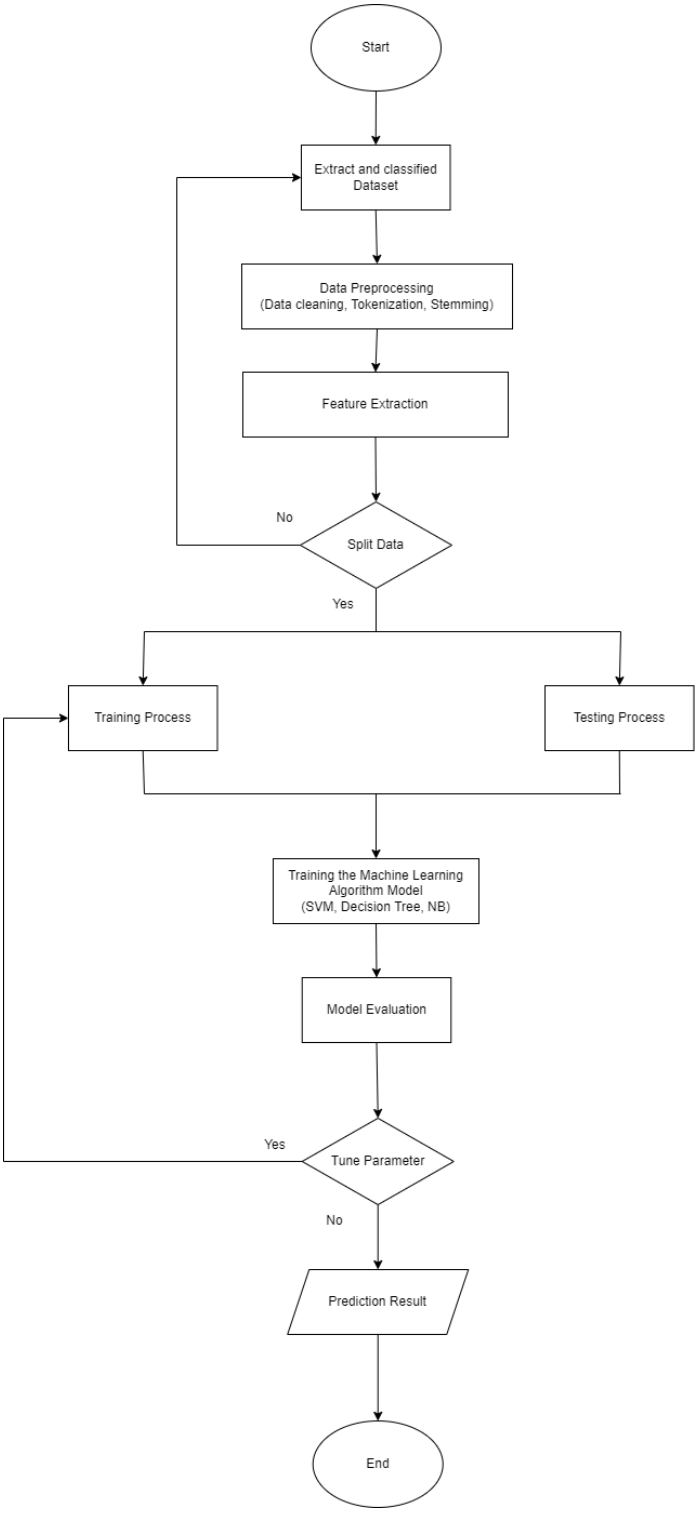


Figure 3.9 Flowchart

3.8.2 Flowchart Explanation

The process starts with data collection and classification. The data is collected from Reddit website via Kaggle. Besides, data classification will be done. Data set will be partitioned in a 70-30 split where 70% for training and 30% for testing. The training data set is used to ensure that the machine detects data patterns, and data cross-validation is used to ensure that the Machine Learning Algorithm is more accurate and efficient while the testing data set is used to verify the model's accuracy once it has been trained. Parameters and features will be set before data pre-processing.

After that, the data will firstly undergo data pre-processing such as data cleaning, tokenization and stemming. Data Pre-processing is the process of preparing information in order to achieve excellent results. (Agarkhed & Reddy, 2021). In this step, hashtags, links, special characters, and digits will be removed. Emojis also will be decode by transforming them to string (Rezig, 2021). Tokenization is a process that removes any non-English characters, lowercases all posts, converts hash tags to regular words, and removes any extraneous letters, spaces, and other elements.

After the data cleaned, Following the application of a machine learning algorithm to the obtained data, Machine Learning Algorithm models will determine the result. Over time, the models will get better at predicting with training. In short, Machine Learning Algorithm models used in here are Support Vector Machine (SVM), Decision Tree and Naïve Bayes (NB).

Once training is complete, model evaluation is undergone to check it’s performing. It assists in determining the optimal model to describe the data and predicting how well that model will perform in the future. This is done by testing the performance of the model on previously unseen data which is the testing data. If testing is done on the training data, the measure will be not accurate. When applied on testing data, the model's performance and speed may be accurately measured.

After evaluated the model, parameter tuning done to see whether the any aspect of accuracy can be improved. This is accomplished by fine-tuning the model's parameters. Parameters are the variables in the model. For instance, during training, the times flow over the training dataset. As the time goes on, the precision will improve. The

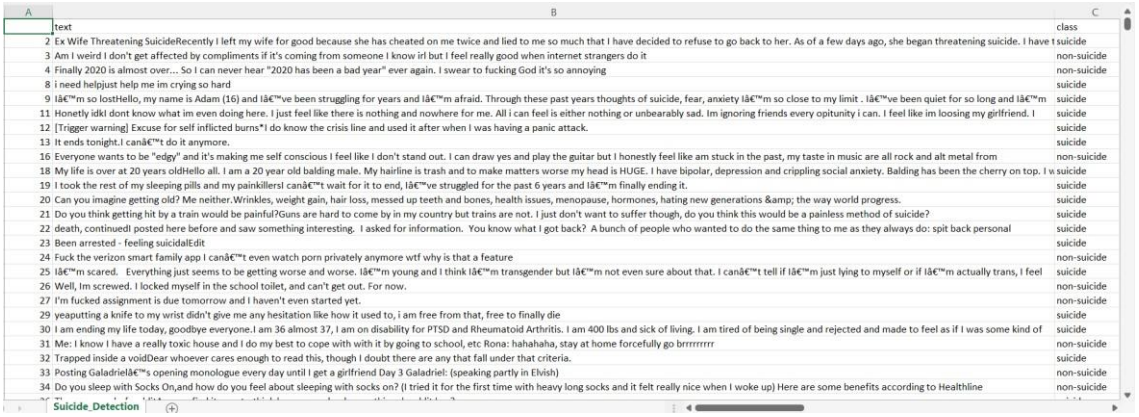
model can be used to create accurate predictions once the training and parameters are satisfied.

3.9 DATA DESIGN

3.9.1 Dataset Description

The dataset collected is a collection of posts from “SuicideWatch” which is subreddits of the Reddit Platform through Kaggle. The Reddit is a social media network built on forums that records the exchange of messages between the author of the original post and the person who commented on it (Shazwani, 2019). Posts are organised into "communities" or "subreddits," which are user-created boards, based on their subject. SuicideWatch is one of the subreddits which is peer support for anyone struggling with suicidal thoughts. All of the postings were made to "SuicideWatch" between December 16, 2008 and January 2, 2021, and were collected using the Pushshift API. The dataset filename is Suicide_Detection.csv. The number of data collected is 233440 and divide into three columns. First is number of data, second is text content and third one is a class which divide into suicidal and non-suicidal. However, only 134438 data will run in this project as the total dataset is too large which generate out of memory error in laptop and consume a long time to run.

Website of dataset: https://www.kaggle.com/datasets/nikhileswarkomati/suicide-watch



	A	B	C
1		text	class
2		Ex Wife Threatening SuicideRecently I left my wife for good because she has cheated on me twice and lied to me so much that I have decided to refuse to go back to her. As of a few days ago, she began threatening suicide. I have t	suicide
3		Am I weird I don't get affected by compliments if it's coming from someone I know irl but I feel really good when internet strangers do it	non-suicide
4		Finally 2020 is almost over... So I can never hear "2020 has been a bad year" ever again. I swear to fucking God it's so annoying	non-suicide
8		i need helpjust help me im crying so hard	suicide
9		Iâ€™m so lostHello, my name is Adam (16) and Iâ€™ve been struggling for years and Iâ€™m afraid. Through these past years thoughts of suicide, fear, anxiety Iâ€™m so close to my limit . Iâ€™ve been quiet for so long and Iâ€™m	suicide
11		Honetly idkI dont know what im even doing here. I just feel like there is nothing and nowhere for me. All i can feel is either nothing or unbearably sad. Im ignoring friends every opitunity I can. I feel like im loosing my girlfriend. I	suicide
12		[Trigger warning] Excuse for self inflicted burns*I do know the crisis line and used it after when I was having a panic attack.	suicide
13		It ends tonight.I canâ€™t do it anymore.	suicide
16		Everyone wants to be "edgy" and it's making me self conscious I feel like I don't stand out. I can draw yes and play the guitar but I honestly feel like am stuck in the past, my taste in music are all rock and alt metal from	non-suicide
18		My life is over at 20 years oldHello all. I am a 20 year old balding male. My hairline is trash and to make matters worse my head is HUGE. I have bipolar, depression and crippling social anxiety. Balding has been the cherry on top. I w	suicide
19		I took the rest of my sleeping pills and my painkillersI canâ€™t wait for it to end, Iâ€™ve struggled for the past 6 years and Iâ€™m finally ending it.	suicide
20		Can you imagine getting old? Me neither.Wrinkles, weight gain, hair loss, messed up teeth and bones, health issues, menopause, hormones, hating new generations & the way world progress.	suicide
21		Do you think getting hit by a train would be painful?Guns are hard to come by in my country but trains are not. I just don't want to suffer though, do you think this would be a painless method of suicide?	suicide
22		death, continuedI posted here before and saw something interesting. I asked for information. You know what I got back? A bunch of people who wanted to do the same thing to me as they always do: spit back personal	suicide
23		Been arrested - feeling suicidalEdit	suicide
24		Fuck the verizon smart family app I canâ€™t even watch porn privately anymore wtf why is that a feature	non-suicide
25		Iâ€™m scared. Everything just seems to be getting worse and worse. Iâ€™m young and I think Iâ€™m transgender but Iâ€™m not even sure about that. I canâ€™t tell if Iâ€™m just lying to myself or if Iâ€™m actually trans, I feel	suicide
26		Well, Im screwed. I locked myself in the school toilet, and can't get out. For now.	non-suicide
27		I'm fucked assignment is due tomorrow and I haven't even started yet.	non-suicide
29		yeaputting a knife to my wrist didn't give me any hesitation like how it used to, i am free from that, free to finally die	suicide
30		I am ending my life today, goodbye everyone.I am 36 almost 37, I am on disability for PTSD and Rheumatoid Arthritis. I am 400 lbs and sick of living. I am tired of being single and rejected and made to feel as if I was some kind of	suicide
31		Me: I know I have a really toxic house and I do my best to cope with with it by going to school, etc Rona: hahahaha, stay at home forcefully go brrrrrrrrr	non-suicide
32		Trapped inside a voidDear whoever cares enough to read this, though I doubt there are any that fall under that criteria.	suicide
33		Posting Galadrielâ€™s opening monologue every day until I get a girlfriend Day 3 Galadriel: (speaking partly in Elvish)	non-suicide
34		Do you sleep with Socks On,and how do you feel about sleeping with socks on? (I tried it for the first time with heavy long socks and it felt really nice when I woke up) Here are some benefits according to Healthline	non-suicide

Figure 3.10 Reddit Dataset from Kaggle

3.10 PROOF OF INITIAL CONCEPT

3.10.1 Machine Learning Algorithm

Three Machine Learning Algorithms will be used in this research to get the results of suicidal ideation prediction which are Support Vector Machine (SVM), Decision Tree and Navies Bayes (NB). These three Supervised Learning Machine Learning Algorithms use label datasets to train algorithms that accurately classify data or predict outcomes. As the dependent variable or result variable in this study, there are two possible outcomes: suicidal or non-suicidal.

3.10.1.1 Support Vector Machine

Support vector machines are supervised machine learning algorithms that attempt to discover the optimum hyperplane in an N-dimensional space that distinguishes between data points, with N denoting the number of features (Rabani et al., 2020). The kernel approach allows SVM to do non-linear classification by mapping data to higher dimensions. Using the kernel, the new data presented to the classifier is also translated to higher dimensions (Kamiş & Goularas, 2019). Hyperplane is the centre line between two distinct values of which have been categorised. It also generates two margin lines with some distance between them so that each categorization point may be easily linearly separated. Margin planes will be linear in hyperplane and pass through the classifier's nearest point. Marginal distance is the distance between two margin planes that aids in the classification of problems. There may be several hyperplanes and marginal distances, but the marginal distance with the greatest margin will be obtained, making categorization easier and more certain (T. Jain et al., 2021).

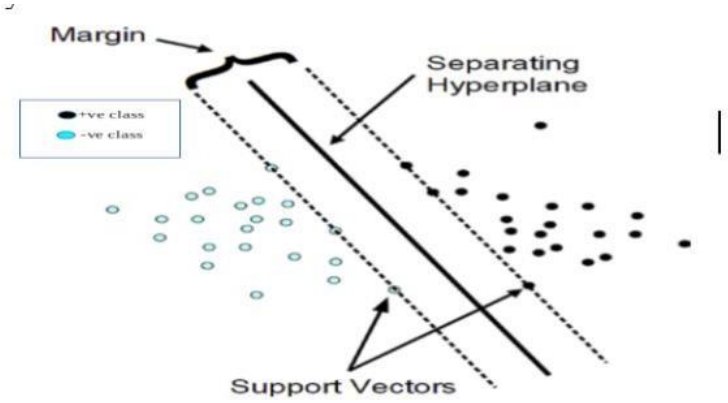


Figure 3.11 Support Vector Machine

In a nutshell, The Support Vector Machine Classifier's purpose is to discover the optimal line or decision boundary for categorising n-dimensional space into classes in the future, so that new data points can be placed in the appropriate category quickly. (Sakib et al., 2021).

3.10.1.2 Decision Tree

Decision Tree is the type of supervised learning algorithms that can be used for both classification and regression problems (Sharma, 2018). It puts the best attribute at the very top of the tree. The training set is partitioned into subsets, with each subset having the same attribute value. Then, on both subsets, the processes above are repeated until leaf nodes are discovered on each branch of the tree (Rabani et al., 2020). The decision tree is drawn upside down, starting from the top, with the root at the top and the leaf at the bottom (Patel & Soni, 2021).

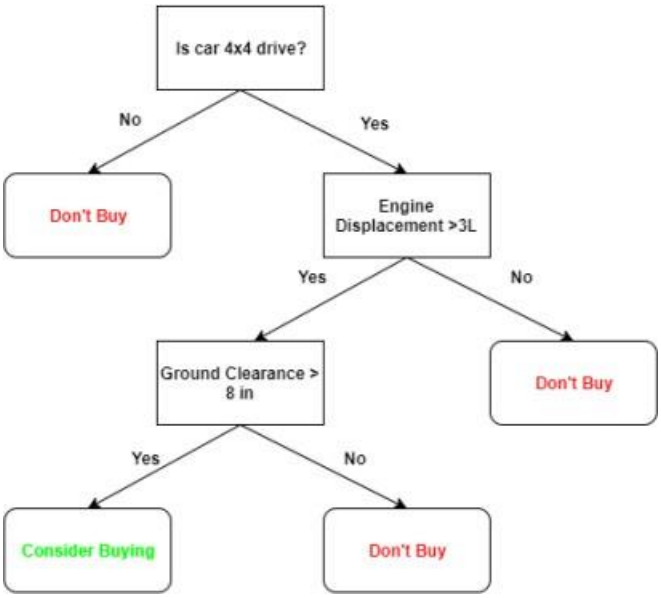


Figure 3.12 Decision Tree

The dependent attribute's outcomes are used to build the decision tree. As a result, entropy must be discovered. The probability of receiving a yes or no, or a 0 or 1 is referred to as entropy. The more entropy there is, the more difficult it is to draw any conclusions. The formula below will be used to find the entropy of the class attribute:

$$E(S) = \sum_{i=1}^{c} -pi\,;log2\ pi$$

Figure 3.13 Formula of Decision Tree

Where E is Entropy, s is set, and Pi is the probability of an event *i* of state S or Percentage of class *i* in a node of state S.

In a nutshell, Decision Tree is used to build a training model that can predict the class or value of a target variable (Sakib et al., 2021). Because the decision tree has a tree-like form, the rationale behind it is simple to comprehend.

3.10.1.3 Naive Bayes

Naïve Bayes (NB) is a Supervised Machine Learning algorithm, input data type is labelled. Naïve Bayes algorithm is based on Bayes algorithm which works on conditional probability (T. Jain et al., 2021). Conditional probability is a measure of the likelihood of an event occurring based on the assumption, supposition, assertion, or evidence of the occurrence of another event (Martfnez-Arroyo & Sucar, 2006).

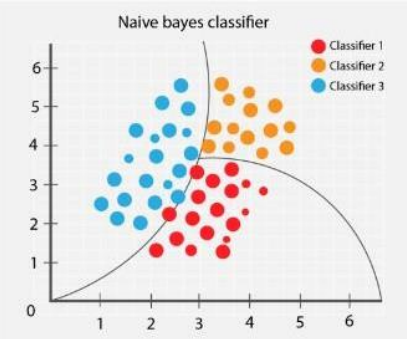


Figure 3.14 Naïve Bayes

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Figure 3.15 Formula of Naïve Bayes

$P(A|B)$ called posterior probability which represent the likelihood of B occurring given proof that A has previously occurred, or how often A occurs when B occurs.

$P(B|A)$ represent the likelihood of B occurring given proof that A has already occurred, or how often B occurs given that A has previously occurred.

$P(A)$ represent how likely A is on its own and the probability of A occurring.

$P(B)$ represent how likely B is on its own and the probability of B occurring.

In short, Naive Bayes is a basic and fast Machine Learning Algorithm for predicting a set of datasets. It's the most popular text classification algorithm, and it can handle multi-class classifications.

3.10.2 Python Library Used

Python machine learning libraries become the most preferred language for Machine Learning Algorithm implementations. Python Machine Learning Libraries used are Numpy, Scikit-learn, Pandas and Matplotlib.

3.10.2.1 Numpy

Numpy is a well-known Python toolkit for processing large multi-dimensional arrays and matrices. It comes in handy for basic scientific computations in Machine Learning. The command for import Numpy is shown in below:

```
import numpy as np
```

Figure 3.16 Numpy Command

3.10.2.2 Scikit-learn

Scikit-learn is a popular Machine Learning library for implementing Machine Learning Algorithms. Numpy and Scipy, two Python libraries, were used to create Scikit-learn. It also works with both supervised and unsupervised techniques. The core function of Scikit-learn in classification, regression, clustering, dimensionality reduction, model selection, and pre-processing.

Scikit-learn used to split dataset into training dataset and testing dataset:

```
# Splitting the dataset into training and test set.
from sklearn.model_selection import train_test_split
```

Scikit-learn used to fit the classifier to the training set such as Decision Tree:

```
#Fitting Decision Tree classifier to the training set
From sklearn.tree import DecisionTreeClassifier
```

3.10.2.3 Pandas

Pandas is a python library for data analysis which is not directly related to Machine Learning. Pandas handle data extraction and preparation as the dataset must be prepared before training. The command used to import Pandas:

```
# importing pandas as pd
import pandas as pd
```

Figure 3.17 Pandas Command

3.10.2.4 Matplotlib

Matplotlib is a python library for data visualization. It also not directly related to Machine Learning. It is an extension of Scipy and able handle Numpy data as well as complex data models made by Pandas. It is handled in a 2D plotting library for making 2D graphs and plots, which is used to visualise the patterns in the data. Matplotlib contains a module named pyplot which provides features to adjust line styles, font settings, and create numerous types of graphs for data display. The command used to import Matplotlib:

```
import matplotlib.pyplot as mtp
```

Figure 3.18 Matplotlib Command

3.11 SOFTWARE EQUIPMENT

Software equipment used in this research as shown in the table below:

Software and its specification	Description
Microsoft 365 Word	Used to do for report writing documents
Microsoft 365 PowerPoint	Used to do the presentation slide
Microsoft 365 Excel	Used to store the dataset with format .csv file
Draw.io	Used to draw flow chart
Google Chrome Version 102.0.5005.63	Used to search the research paper and information for the project
Google Colaboratory	Used to write and run the python code
Visual Studio Code	Used to run the prototype 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

3.12 HARDWARE EQUIPMENT

The hardware used is effective and compatible:

Hardware	Specification	Description
Laptop	HP Pavilion Laptop 15 CPU: Intel(R) Core(TM) i5-8265U CPU @ 1.60GHz2 1.80 GHz VGA: NVIDIA GeForce MX250 VRAM: 128 MB	To perform all the documentation and development.
Smartphone	Iphone XR CPU: Hexa-core (2x2.5 GHz Vortex + 4x1.6 GHz Tempest) GPU: Apple GPU (4-core graphics) OS: iOS 15.5	Used to aid in search for information needed to complete the research

Table 3.2 Hardware equipment and its specification

3.13 POTENTIAL USE OF PROPOSED SOLUTION

In future, I hope to understand how analysis of social media behaviour can lead to predict suicidal ideation. For future work, the same dataset also can be further enhanced by using several others different Machine Learning Algorithms such as Random Forest, Logistic Regression, and many more to compare the performance and accuracy of each classifier. Besides, the three Machine Learning Algorithms used in this research, Support Vector Machine (SVM), Decision Tree and Naïve Bayes (NB) can implement on different dataset to do the prediction result about suicidal ideation via social media. They also can predict not only suicidal or non-suicidal but also predict other mental such as depression, anxiety and others.

CHAPTER 4

RESULT AND DISCUSSION

4.1 INTRODUCTION

This chapter discuss and explain the outcomes on the development that have been implement using the technique, method and equipment that suitable for this project. Lastly in the result and discussion section, it shows the result that supposedly fulfil the project objectives.

4.2 RESULT

Reddit dataset is collected from Kaggle websites which collected from 2008 until 2021. After proper pre-processing the data were labelled as “suicide” and “non-suicide” in other words as 1 and 0. The total dataset from Kaggle is 233440 data but only 134438 data will run in this project as the total dataset is too large which generate out of memory error in laptop and consume a long time to run. A dataset consisting of 67282 non-suicide data and 67156 suicide data was used. Python was used as language to implement the work. Besides, python machine learning libraries such as NumPy, Pandas, NLTK, Matplotlib are used.

The dataset was undergoing pre-processed to remove stop words, numbers, special characters and convert all uppercase letter to lowercase letter. Word Cloud is import for data visualization for all words, positive words and negative words. The more often and how important a word is mentioned in a document, the bigger and bolder it appears.

After that, data is split into 70% for training the model which is 94106 data and 30% for testing the model which equal to 40332 data. Three machine learning algorithms: Support Vector Machine (SVM), Naïve Bayes and Decision Tree were applied to evaluate the performance for train set and test set and compared between them. Output accuracy, precision, recall and F1_score of the three algorithms will be evaluated to get the most accurate algorithm. Results will show in graph and table. Table below shows the output of these implemented machine learning algorithms.

Furthermore, Confusion Matrix will be generated for training and testing data of the most accurate machine learning algorithm. It is used to represent the true positive, true negative, false positive and false negative values.

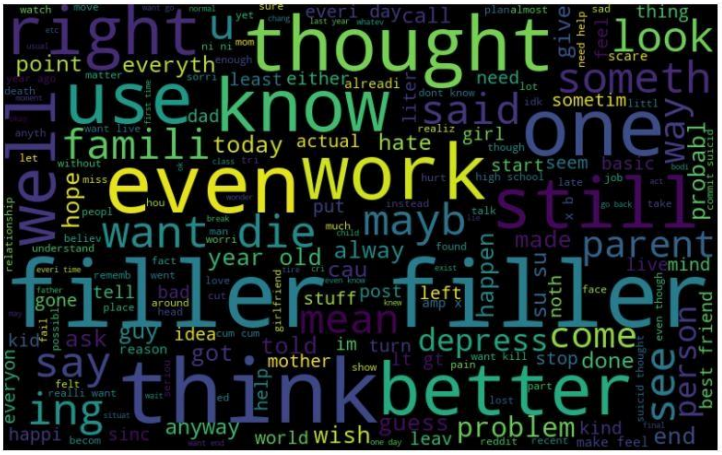


Figure 4.1: All Words in Word Cloud

Figure 4.2: Positive Words in Word Cloud

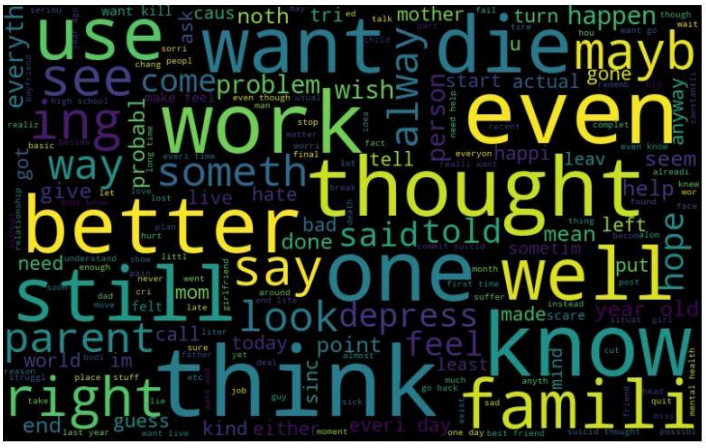


Figure 4.3: Negative Words in Word Cloud

Model	Train				Test			
	Accuracy	**Precision**	**Recall**	**F1_Score**	**Accuracy**	**Precision**	**Recall**	**F1_Score**
Support Vector Machines	0.9829	0.9814	0.9844	0.9829	0.9340	0.9451	0.9214	0.9331
Naïve Bayes	0.8867	0.8280	0.9759	0.8959	0.8745	0.8148	0.9690	0.8852
Decision Tree	0.9510	0.9118	0.9893	0.9490	0.8579	0.8882	0.8186	0.8520

Table 4.1: Accuracy, Precision, Recall and F1_Score of Machine Learning Models

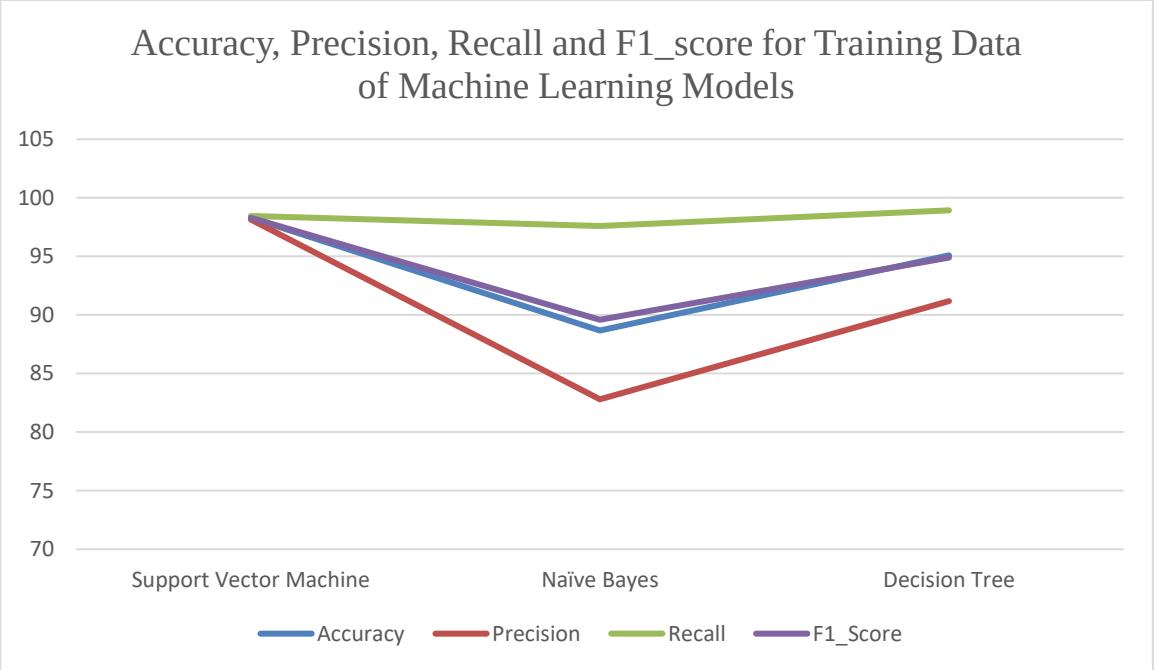


Figure 4.4: Accuracy, Precision, Recall and F1_score for Training Data of Machine Learning Models

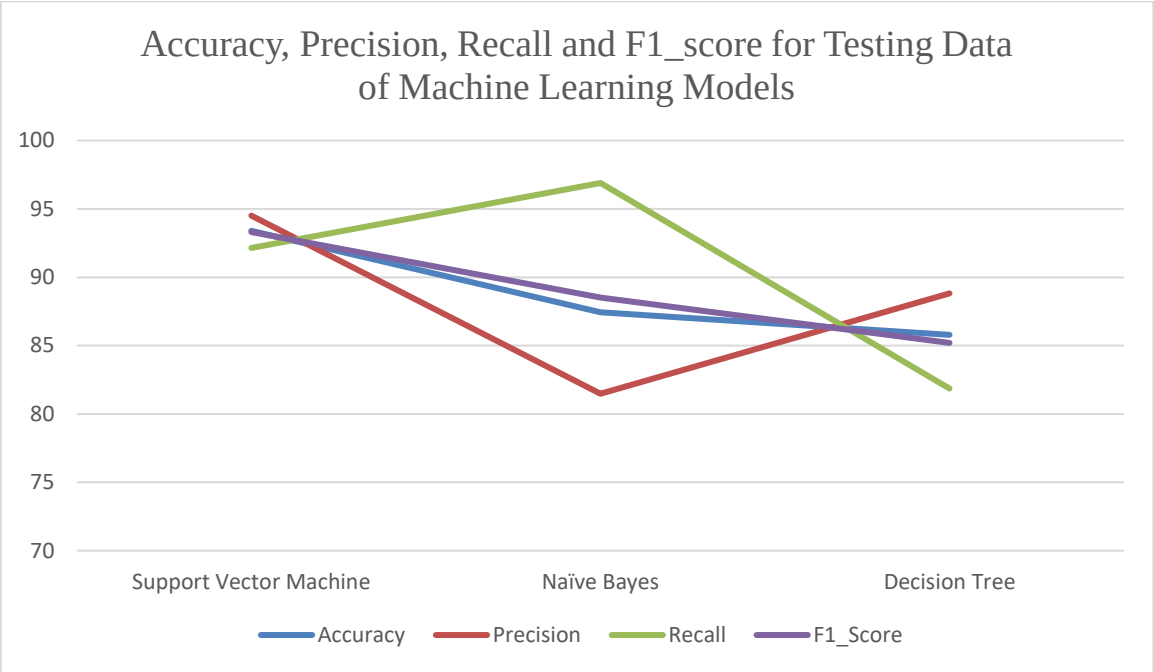


Figure 4.5: Accuracy, Precision, Recall and F1_score for Testing Data of Machine Learning Models

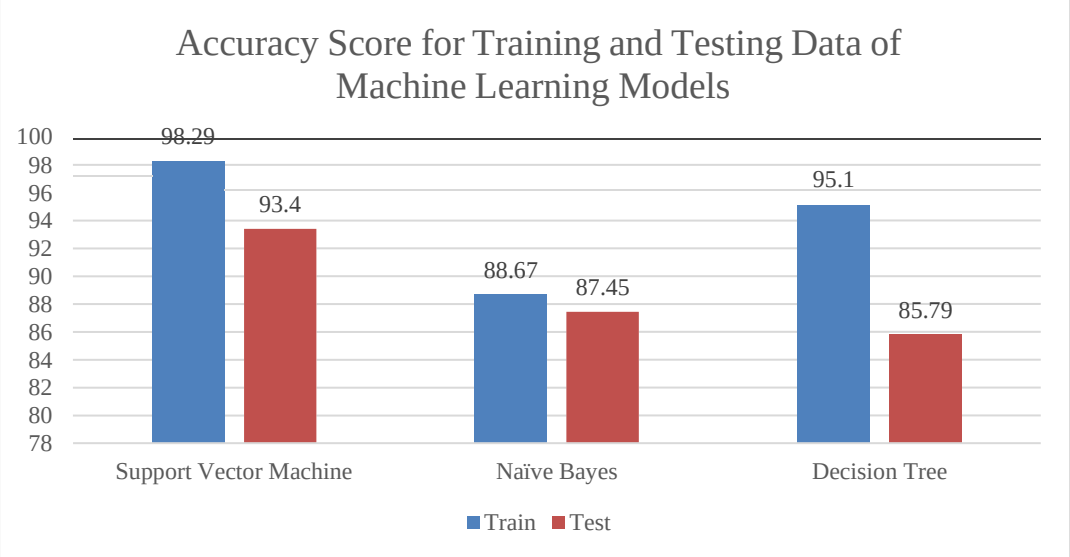


Figure 4.6: Accuracy Score of Machine Learning Models

From figure above, it shows that the accuracy score of SVM, Naïve Bayes and Decision Tree of train data are 98.29%, 88.67% and 95.10% accordingly and for test data are 93.40%, 87.45% and 85.79% accordingly.

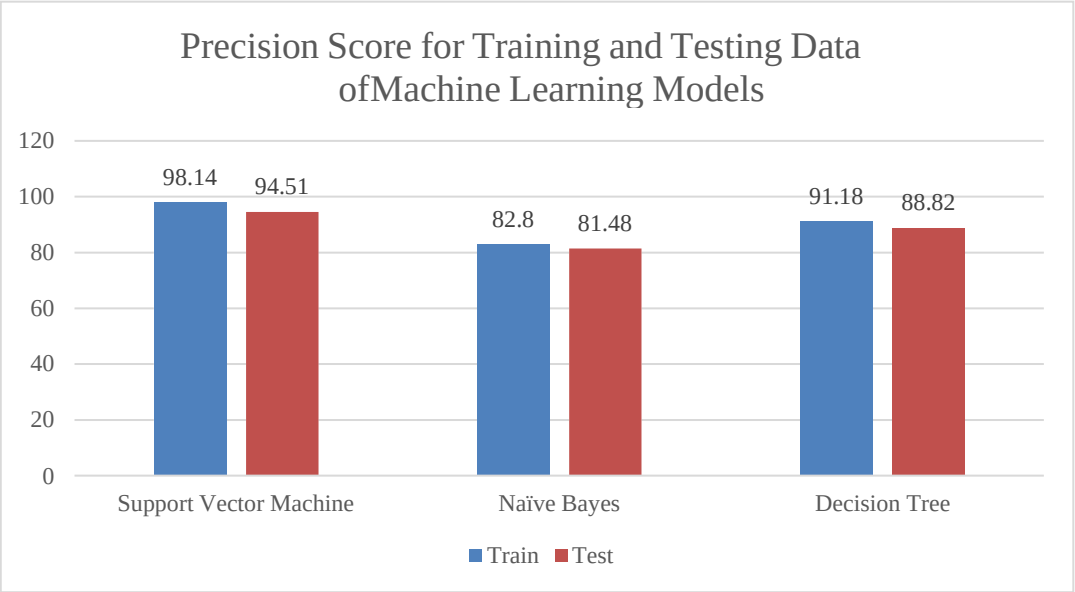


Figure 4.7: Precision Score of Machine Learning Models

The precision score of train data for SVM is 98.14%, for Naïve Bayes is 82.80% and for Decision Tree is 91.18% while of test data for SVM is 94.51%, for Naïve Bayes is 81.48% and for Decision Tree is 88.82%.

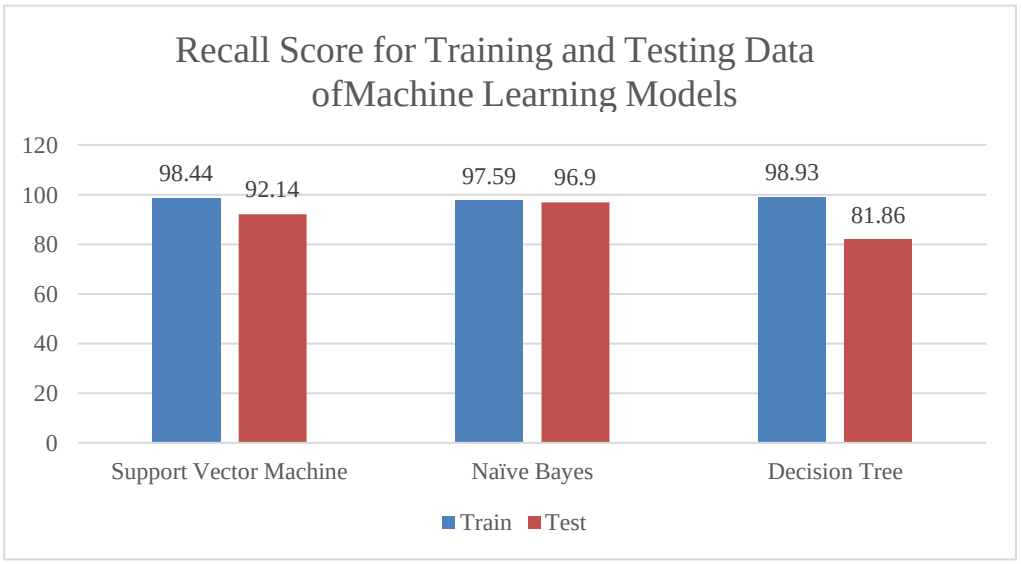


Figure 4.8: Recall Score of Machine Learning Models

The recall score of SVM, Naïve Bayes and Decision Tree of train data are 98.44%, 97.59% and 98.93% accordingly and for test data are 92.14%, 96.90% and 81.86% accordingly.

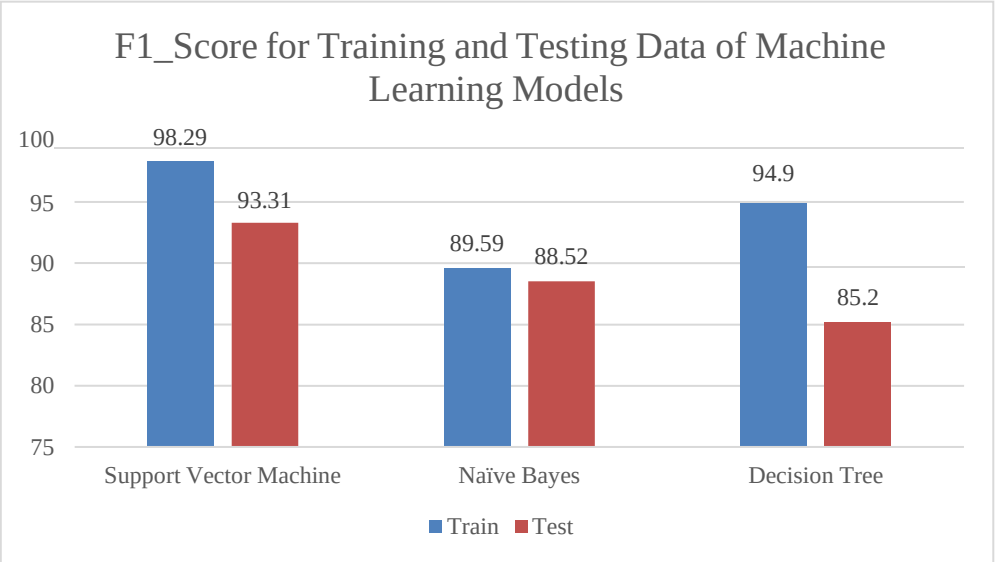


Figure 4.9: F1_Score of Machine Learning Models

F1_Score of SVM, Naïve Bayes and Decision Tree of train data are 98.29%, 89.59% and 94.90% accordingly and for test data are 93.31%, 88.52% and 85.20% accordingly.

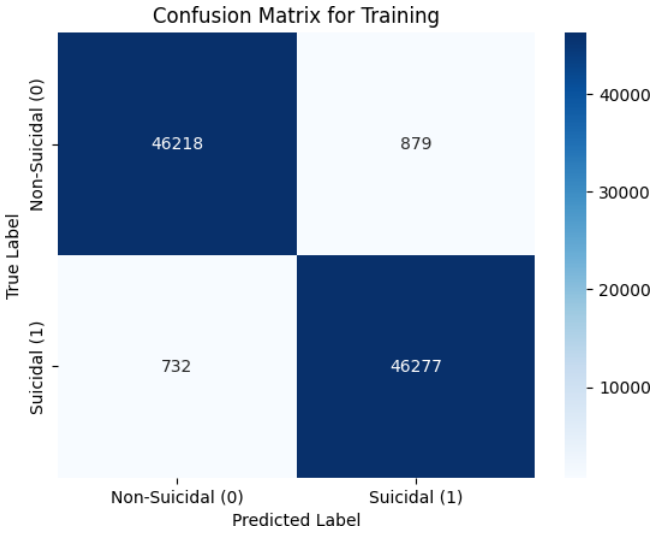


Figure 4.10: Confusion Matrix for SVM for training data

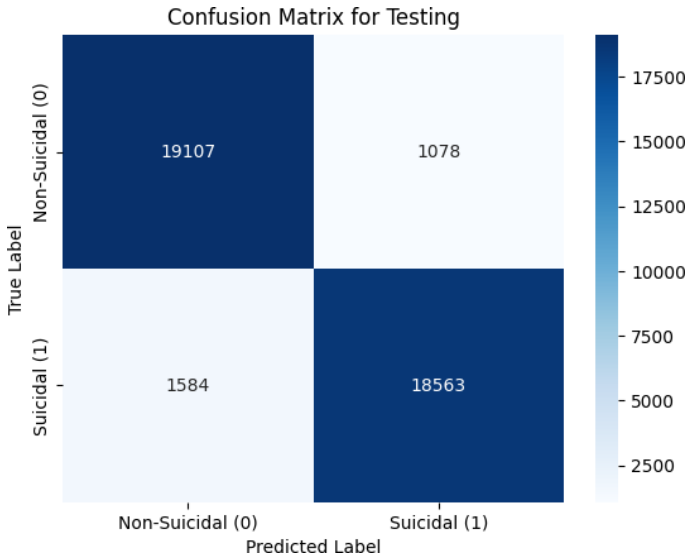


Figure 4.11: Confusion Matrix for SVM for testing data

In short, based on accuracy, SVM algorithm is the most accurate for training data which is 98.29% compared to Naïve Bayes (88.67%) and Decision Tree (95.10%). While SVM also is the most accurate algorithm for testing data which is 93.40% compared to Naïve Bayes (87.45%) and Decision Tree (85.79%). Therefore, Confusion Matrix for training and testing data are generated for SVM.

From the figure 4.11, it shows that 19107 correct results out of 20185 of non-suicidal data was achieved. 1584 correct suicide prediction out of 20147 suicidal data was achieved from the testing data of SVM.

4.3 DISCUSSION

In conclusion, from the result according to the accuracy score for both training and testing data, Support Vector Machines SVM get precise result (98.29%) for the training data and also get the most accuracy result for the testing data (93.40%).

CHAPTER 5

CONCLUSION

5.1 INTRODUCTION

This research is focus on predicting suicidal ideation via social media. It helps to detect suicide attempt from the post before that person end his life. This can reduce the suicideprobability because people can take action before letting it happen. Machine Learning Algorithms implement for the prediction of suicidal ideation via data set from Subredditof Reddit. The dataset undergoes process of feature extraction and data pre-processing before modelling. The models were trained and tested using accuracy score, precision, recall and F1-Score. The result shows that for training data, SVM algorithm give the most accuracy which are 98.29% compared to Naïve Bayes (88.67%) and Decision Tree (95.10%). In testing process, SVM give the precise result with 93.40% compared to NaïveBayes (87.45%) and Decision Tree (85.79%). As conclude, Machine Learning Algorithm can help to predict the post from tweets before suicide case happens.

5.2 LIMITATION

In collecting data, analysing error factors, and predicting results, it was strictly attempted to be as accurate as possible. However, there were some unavoidable limitations. If the dataset is labelled by psychiatrists, then tweet labelling could provide greater accuracy. Besides, when the datasets larger such from Reddit, the accuracy of the model may decrease.

5.3 FUTURE WORK

For future work, the same dataset also can be further enhanced by using several others different Machine Learning Algorithms such as Random Forest, Logistic Regression, and many more to compare the performance and accuracy of each classifier. Besides, the three Machine Learning Algorithms used in this research, Support Vector Machine (SVM), Decision Tree and Naïve Bayes (NB) can implement on different

dataset to do the prediction result about suicidal ideation via social media. They also can predict not only suicidal or non-suicidal but also predict other mental such as depression, anxiety and others.

REFERENCES

Agarkhed, J., & Reddy, M. (2021). Prediction Model for Preventing Suicide Attempts Using Machine Learning. *2021 4th International Conference on Electrical, Computer and Communication Technologies, ICECCT 2021*. https://doi.org/10.1109/ICECCT52121.2021.9616815

Aldhyani, T. H. H., Alsubari, S. N., Alshebami, A. S., Alkahtani, H., & Ahmed, Z. A. T. (2022). Detecting and Analyzing Suicidal Ideation on Social Media Using Deep Learning and Machine Learning Models. *International Journal of Environmental Research and Public Health*, *19*(19). https://doi.org/10.3390/ijerph191912635

Braithwaite, S. R., Giraud-Carrier, C., West, J., Barnes, M. D., & Hanson, C. L. (2016). Validating machine learning algorithms for twitter data against established measures of suicidality. *JMIR Mental Health*, *3*(2), 1–10. https://doi.org/10.2196/mental.4822

De Choudhury, M., & Kiciman, E. (2017). The language of social support in social media and its effect on suicidal ideation risk. *Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017*, *Icwsm*, 32–41. https://doi.org/10.1609/icwsm.v11i1.14891

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

Jain, S., Narayan, S. P., Dewang, R. K., Bhartiya, U., Meena, N., & Kumar, V. (2019). A Machine Learning based Depression Analysis and Suicidal Ideation Detection System using Questionnaires and Twitter. *2019 IEEE Students Conference on Engineering and Systems, SCES 2019*, 1–6. https://doi.org/10.1109/SCES46477.2019.8977211

Jain, T., Jain, A., Hada, P. S., Kumar, H., Verma, V. K., & Patni, A. (2021). Machine Learning Techniques for Prediction of Mental Health. *Proceedings of the 3rd International Conference on Inventive Research in Computing Applications, ICIRCA 2021*, 1606–1613. https://doi.org/10.1109/ICIRCA51532.2021.9545061

Kamiş, S., & Goularas, D. (2019). Evaluation of Deep Learning Techniques in Sentiment Analysis from Twitter Data. *Proceedings - 2019 International Conference on Deep Learning and Machine Learning in Emerging Applications, Deep-ML 2019*, 12–17. https://doi.org/10.1109/Deep-ML.2019.00011

Martfnez-Arroyo, M., & Sucar, L. E. (2006). Learning an optimal naive Bayes classifier. *Proceedings - International Conference on Pattern Recognition*, *3*, 1236–1239. https://doi.org/10.1109/ICPR.2006.748

Mbarek, A., Jamoussi, S., Charfi, A., & Ben Hamadou, A. (2019). Suicidal profiles detection in twitter. *WEBIST 2019 - Proceedings of the 15th International Conference on Web Information Systems and Technologies*, *Webist*, 289–296. https://doi.org/10.5220/0008167602890296

Patel, H., & Soni, N. (2021). Machine Learning Based Approach for Prediction of Suicide Related Activity. *Proceedings - 2nd International Conference on Smart Electronics and Communication, ICOSEC 2021*, 967–972. https://doi.org/10.1109/ICOSEC51865.2021.9591836

Poulin, C., Shiner, B., Thompson, P., Vepstas, L., Young-Xu, Y., Goertzel, B., Watts, B., Flashman, L., & McAllister, T. (2014). Predicting the risk of suicide by analyzing the text of clinical notes. *PLoS ONE*, *9*(1), 1–7. https://doi.org/10.1371/journal.pone.0085733

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

Rabani, S. T., Rayees Khan, Q., & Ud Din Khanday, A. M. (2020). Multi-Class Suicide Risk Prediction on Twitter Using Machine Learning Techniques. *Proceedings - IEEE 2020 2nd International Conference on Advances in Computing, Communication Control and Networking, ICACCCN 2020*, 128–134. https://doi.org/10.1109/ICACCCN51052.2020.9362979

Rezig, A. A. (2021). A Novel Optimizer Technique for Suicide Prediction in Twitter Environment. *Proceedings - 2021 International Conference on Information Systems and Advanced Technologies, ICISAT 2021*. https://doi.org/10.1109/ICISAT54145.2021.9678419

Ribeiro, J. D., Pease, J. L., Gutierrez, P. M., Silva, C., Bernert, R. A., Rudd, M. D., & Joiner, T. E. (2012). Sleep problems outperform depression and hopelessness as cross-sectional and longitudinal predictors of suicidal ideation and behavior in young adults in the military. *Journal of Affective Disorders*, *136*(3), 743–750. https://doi.org/10.1016/j.jad.2011.09.049

Sakib, T. H., Ishak, M., Jhumu, F. F., & Ali, M. A. (2021). Analysis of suicidal tweets from twitter using ensemble machine learning methods. *2021 International Conference on Automation, Control and Mechatronics for Industry 4.0, ACMI 2021*, *0*(978), 8–9. https://doi.org/10.1109/ACMI53878.2021.9528252

Sharma, S. (2018). Predictive analysis using classification techniques in healthcare domain. *International Journal of Computer Sciences and Engineering*, *6*(2), 206–212. https://doi.org/10.26438/ijcse/v6i2.206212

Shazwani. (2019). A Study of User Behaviors and Activities on Online Mental Health Communities. *Αγαη*, *8*(5), 55.

Vioules, M. J., Moulahi, B., Aze, J., & Bringay, S. (2018). Detection of suicide-related posts in Twitter data streams. *IBM Journal of Research and Development*, *62*(1). https://doi.org/10.1147/JRD.2017.2768678

Yeskuatov, E., Chua, S. L., & Foo, L. K. (2022). Leveraging Reddit for Suicidal Ideation Detection: A Review of Machine Learning and Natural Language Processing Techniques. *International Journal of Environmental Research and Public Health*, *19*(16). https://doi.org/10.3390/ijerph191610347

APPENDIX A

Gantt Chart

Task Name	Start Date	End Date
Idea Proposing	02/13/22	02/26/22
Problem Statements	02/28/22	03/05/22
Objectives and Scope	03/05/22	03/11/22
Literature Review	03/12/22	03/23/22
Prior Research Work	03/24/22	04/13/22
Research Framework	04/14/22	05/14/22
Project Requirement	05/15/22	05/26/22
Proposed Design	05/27/22	05/30/22
Data Design	05/30/22	05/31/22
Proof of Initial Concept	06/01/22	06/02/22
Potential Use of Proposed Solution	06/02/22	06/03/22
Complete Thesis Report	06/01/22	06/12/22
Presentation for PSM 1	06/13/22	06/16/22
Revise and Finalize	06/16/22	06/24/22
Data Preprocessing	06/27/22	07/11/22
Implementation of Machine Learning Models	07/13/22	08/22/22
Training and Testing Process	08/22/22	09/26/22
Result	09/28/22	10/14/22
Discussion	10/17/22	10/31/22
Conclusion	11/02/22	11/07/22
Limitation	11/02/22	11/07/22
Future Work	11/02/22	11/07/22
Presentation PSM	01/31/23	02/02/23
Finalization and Submission PSM	02/24/23	02/24/23

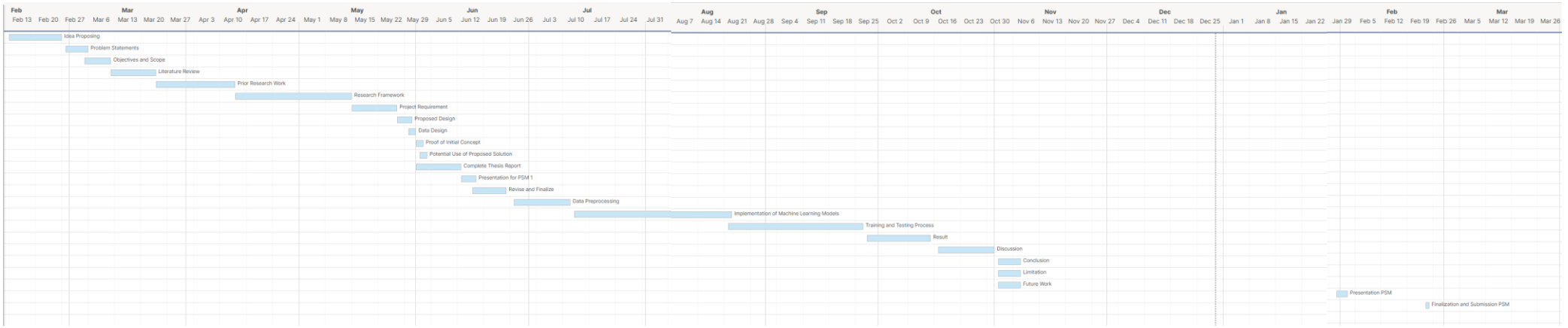


Figure 3.19 Gantt Chart for PSM