DETECTION OF SUICIDAL IDEATION ON SOCIAL MEDIA PROFILES USING MACHINE LEARNING TECHNIQUES

BY

SABBIR HOSSAIN

ID: 211-25-930

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Masters of Science in Computer Science and Engineering

Supervised By

Dr. Touhid Bhuiyan
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

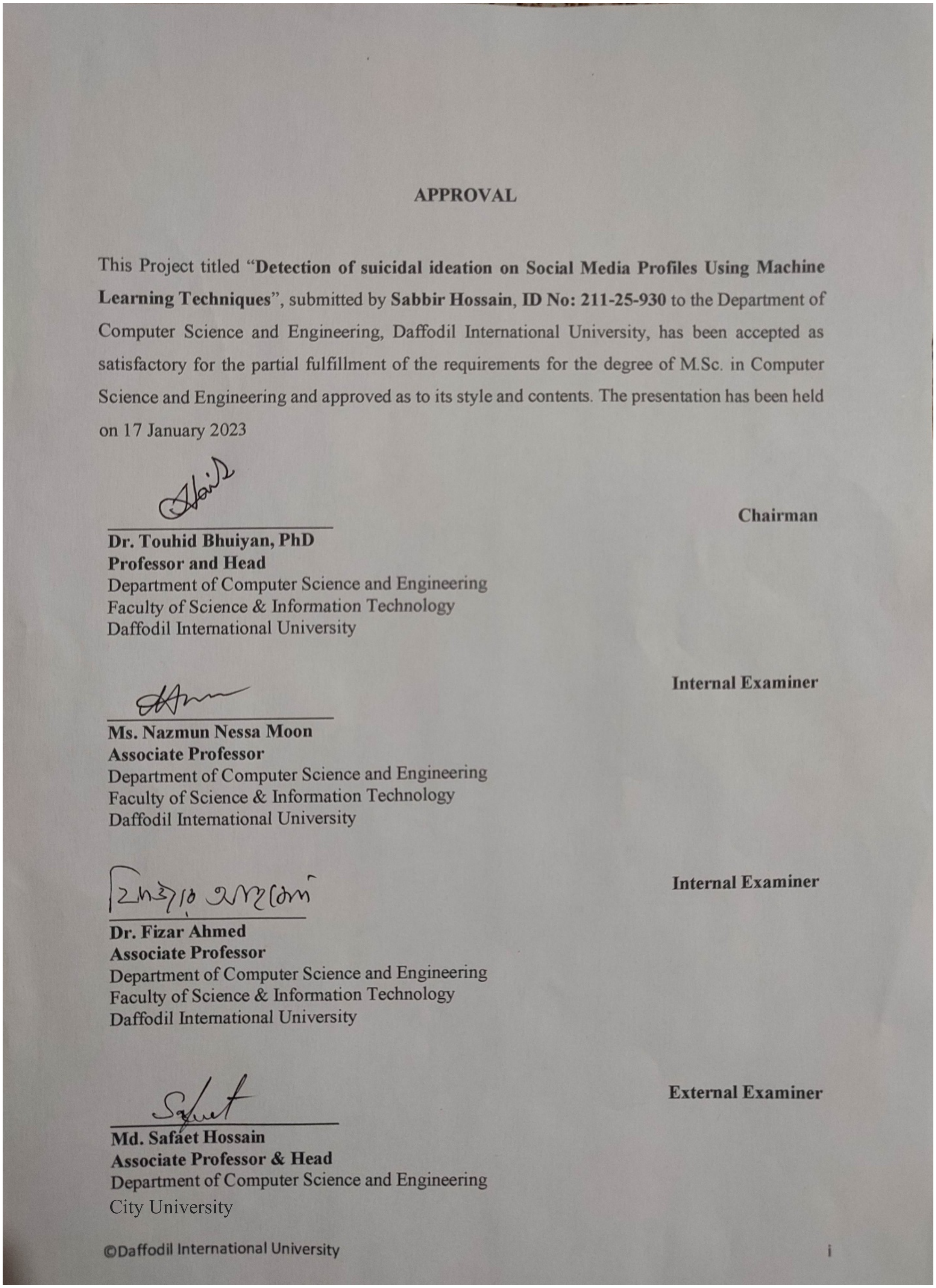
Co-Supervised By

Mr. Abdus Sattar
Assistant Professor & Coordinator M.Sc
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

17 JANUARY 202



APPROVAL

This Project titled "**Detection of suicidal ideation on Social Media Profiles Using Machine Learning Techniques**", submitted by **Sabbir Hossain**, **ID No: 211-25-930** to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of M.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 17 January 2023

Chairman

Dr. Touhid Bhuiyan, PhD
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

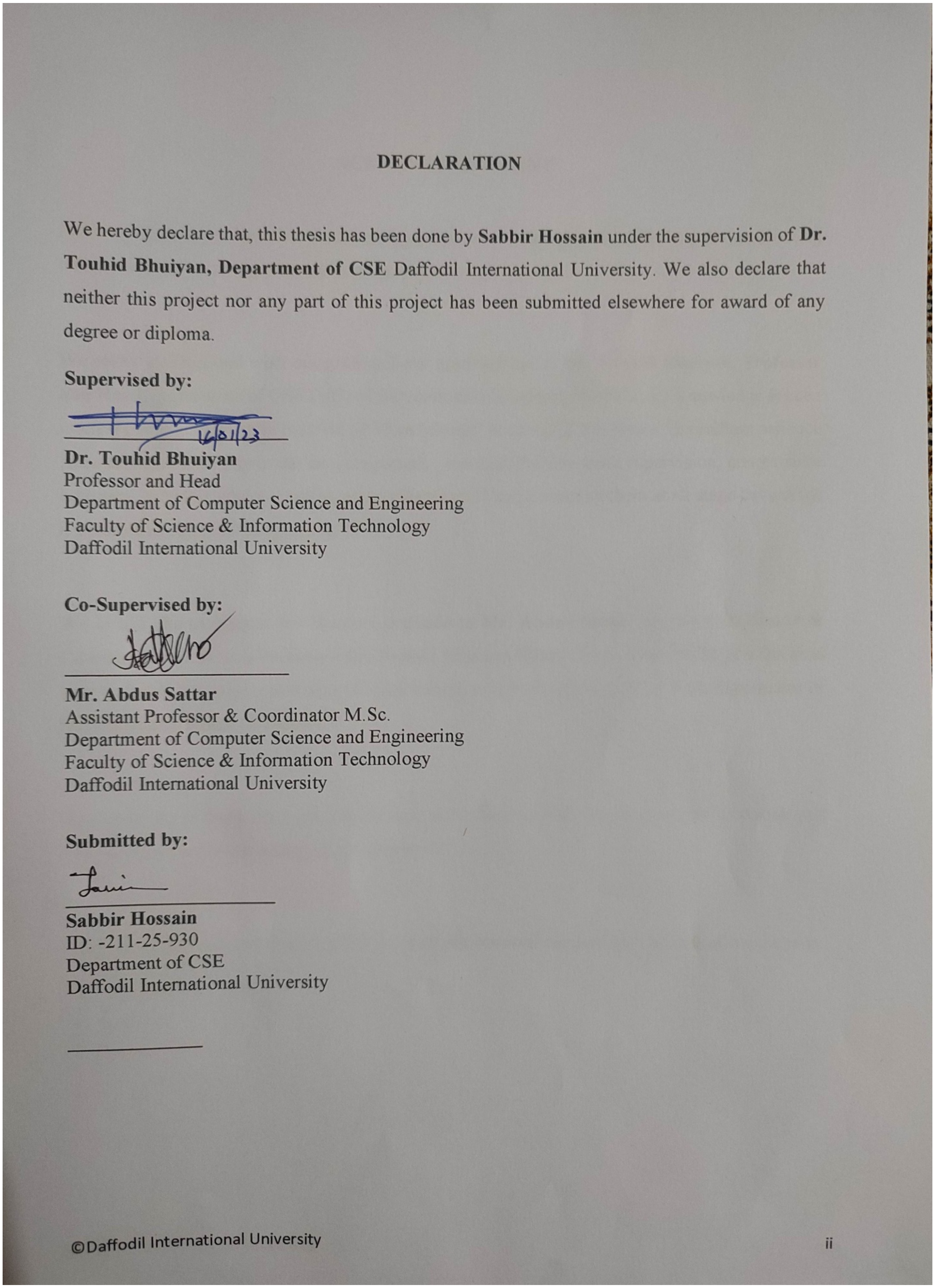
Ms. Nazmun Nessa Moon
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Dr. Fizar Ahmed
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

External Examiner

Md. Safaet Hossain
Associate Professor & Head
Department of Computer Science and Engineering
City University



DECLARATION

We hereby declare that, this thesis has been done by **Sabbir Hossain** under the supervision of **Dr. Touhid Bhuiyan, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

16/01/23

Dr. Touhid Bhuiyan
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Co-Supervised by:

Mr. Abdus Sattar
Assistant Professor & Coordinator M.Sc.
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Submitted by:

Sabbir Hossain
ID: -211-25-930
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to almighty Allah for His divine blessing makes us possible to complete the final year thesis successfully.

We really grateful and wish our profound our indebtedness to **Dr. Touhid Bhuiyan**, **Professor and Head**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "Data Mining" to carry out this thesis. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to **Mr. Abdus Sattar Assistant Professor & Coordinator M.Sc.** and **Professor Dr. Touhid Bhuiyan Head**, Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

Nowadays, due to numerous reasons like the nuclear family, peer pressure for fake prestige, impatience mindset, and mental pressure has become a usual trait in every person. When someone finishes their own life, we say that they passed by suicide. A suicide attempt denotes that someone tried to end their life, but did not die. It has been a crucial issue in current society. For this earlier detection and prevention of suicide attempts should be handled to protect people's life. Today with the expansion of technology people tend to express their emotions on social media. A massive amount of people vents out their emotions online as they have no support system in actual life. It has been noticed or seen a lot of times those suicidal trends varying from mild to an extreme could be from a person's online profile activity. In this article, we put ourselves in a tough context, on the opinions that could be thinking of suicide. Particularly, we propose to address the shortage of terminological resources connected to suicide by a technique of assembling a vocabulary associated with suicide. After that, we proposed a specific method that includes all critical criteria which could be demonstrated by a suicidal person using Natural Language Processing (NLP) techniques. Our approach indicates efficiently an actual, mentally worried profile from a typical profile. Finally, we summarized to encourage future research. We also summarized the limitations of existing work and provide an outlook of further research approaches. This study provides an explanation as well as a solution by classifying the Reddit suicide and non-suicide opinion using various algorithm. Among these algorithms, Logistic Regression accuracy is the best accuracy that is 92.97%. The proposed model is made on Jupyter Notebook (a Python-based IDE) and trained on Kaggle's standard Suicide and depression dataset which has 2,33,338 records.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

1.1 Background of the Research

In this technology-based modern world, powerful social media have altered the way in where a maximum of people express their ideas and real-life thoughts and concern. This chance is provided via text-based journals, in the online conversation areas, development evaluation websites, etc. People depend on this user-generated content heavily. As a result, according to the (WHO), has been said that the suicide rate has been overstated worldwide very rapidly. As people vastly express their opinions on social media, we can take advantage of them. We acknowledge that earlier detection of suicidal behaviors on social media may help prevent deaths. Using machine learning techniques, we can specify necessary steps that can be taken to save a valuable life. Machine learning is being utilized in a variety of fields to solve complex challenges. In social media, there could be less formal information but Machine learning could be utilized to overcome complicated challenges and risk factors. We also can classify low-risk suicide attempters and high-risk ones by this.

1.2 Problem Statement

First, to the finest of our learning, but in we see there is a large number of shortages of proper suicidal research on today's popular social media. In the current method, the technique is made clear to understand the social network and the similar attributes of Reddit content creators whose responsibility is to produce the content placed by a human expert. And it is the process of valid self-considering to process the content, usually, indicated as self-destroying creativity or ideation. In this study is for catching people's sentiments accurately and taking necessary steps to reduce death attempts.

1.3 Aim of the Research

To estimate the individual and integrated capabilities of a very realistic machine learning instance where we can predict the suicide attempts with the help of social media data. This paper aims to study and research suicidal attempts more conveniently and stimulate debate between the various constituencies involved in these fields. Understanding people's concerns early from social media find out the actual risk factors so that we can rescue a people from die. That is why we have chosen to study early suicide detection so that, this research can propose a model to predict these problems early with the highest accuracy rate for contributing to myself, the society and social media managers as well as medical sectors.

1.4 Propose Solution

The systematic model is based on data mining techniques applying machine learning algorithms that can easily classify suicide risk.

1.5 Research Methodology

For development purposes, I used python programming language which is executed by Jupyter Notebook. It is an open-source online instrument that allows a user, scientist, thinker, or analyst to produce and shares a document named Notebook, which includes live programs, documentation, charts, plots, and visualizations. For the dataset, I have used the Reddit Suicide dataset from Kaggle having 232074 Reddit records with the target Suicide and Depression Detection. I have also used a bunch of very effective machine learning algorithms which are called LR, DT, RF, SVM and the KNN Classifiers as classifying techniques

1.5.1 Suicide Detection

Suicide is a crucial subject in current society. In this time of nuclear family structure and addiction to virtuality people face loneliness. The early detection of ideation suicidal attempts and as well as the prevention of the suicide attempts should be managed for protecting somebody's valuable life. According to the information of the (WHO), recently every year more than 703 000 people are taking their life and there are a lot of people who trying attempt suicide. Suicide occurs mainly in the life cycle of in between Fifteen- to Nineteen-year-olds people globally in 2019.

1.5.2 Data Mining

The data mining [17] is a fundamental an initial process of storing relevant information, identifying attributes, looking for fundamental links between attributes, recognizing patterns, and extracting data for socio-commercial objectives. Massive quantities of data are organized in today's digital planet, and this data is studied to define the association between factors, their relevancy to a specific problem, the trends and patterns in which it is developed, and is then utilized by different Data Mining Tools to predict useful and significant outcomes in various areas. Data mining is an essential element for powerful analytics ambitions in associations. The data it generates can be utilized in corporation intellect and advanced analytics applications that concern analysis of climate data, as well as real-time analytics applications that analyze streaming data as it's constructed or collected.

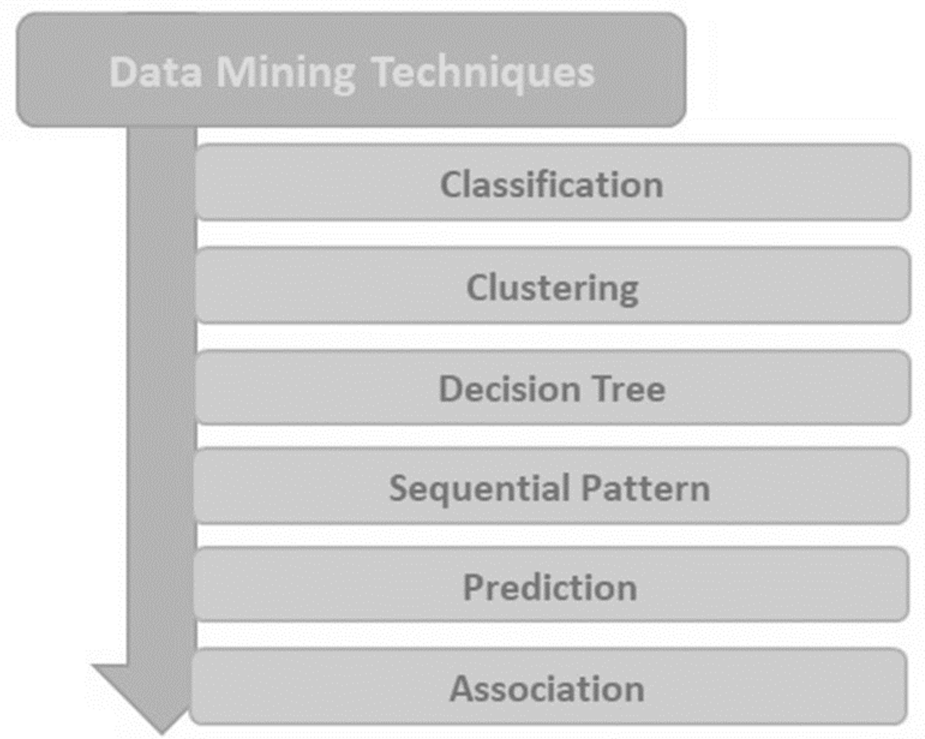


Figure 1.5.2.1: Data mining techniques [26]

Data mining, as a valuable resource for climate change researchers, has played a vital role in research and might potentially be helpful to our suicide understanding. It shows an early suicide risk prediction using social media data.

1.5.3 Machine Learning

The machine learning algorithms are utilized in climate change studies in a systematic method. Not only is the climate modification sector but also predicts future climate statistics. A machine learning algorithm utilize data mining approaches to construct the prototype and pattern to find the exactness rate of classification, prognosis, connection, and many others. The ML is a progressive analysis of mathematically demonstrated. ML is also a very much important element of this field of data science. On the other hand, the machine learning algorithms, construct a mathematical pattern based on example data, which is directed to as "training data," to make forecasts or decisions rather than being closely programmed to carry out the work. [11]

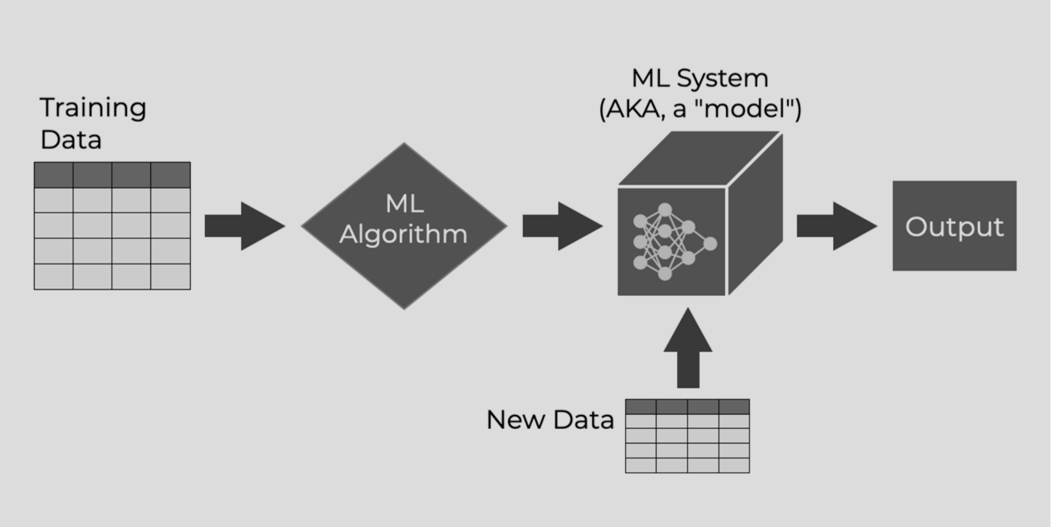


Figure 1.5.3.1 Machine learning techniques [25]

Machine Learning is a very essential area for the artificial intelligence in which computers are instructed to learn on their own by telling them a variety of models, data sets, and patterns that assist in the development.

1.5 Conclusion

After applying 4-5 algorithms I see in this research, Logistic Regression Algorithm got the highest accuracy of 93.00% in analyzing climate changes. The main objective is to increase the efficiency in the prediction rate of early suicide detection sentiment applying correctly the data mining techniques.

CHAPTER 2

Literature Review

2.1 Introduction

The majority of researchers in the data mining field can identify early suicide risk using a variety of machine learning algorithms and approaches, which provides them with a new area of research. There have been many studies on sentiment-based classification in recent years. Machine learning and semantic orientation methodologies can be used to categorize the methods used by researchers. The machine learning approach involves training a classifier using a set of representative data, making it a supervised work. On the other hand, the semantic orientation approach is an unsupervised method that infers the viewpoint of the document as a whole from the semantic orientation of the words that make up the document.

2.2 Related Works

By the efficient use of Machine Learning and also with the assist of Natural language processing technique (NLP), Sravanth, T., Hema, V. and Reddy, S.T. (2020) et al. [1] develop a quantitative measure of standing with Data Mining Methods using various classifiers. In this reacharch they clarified a freshly generated new idea for suicide explanation. After that thay provide an explanation a varification online suicide. They mainly concentrated on social site datas like the tweet, Facebook, and Reddit data. And then they collected suicidal attempts based on data using various tags and keys.

Birjalia, M., Beni-Hssane, A. and Erritali, M. et Al. [2] applied the multiple machine learning technique like (SVM) and also used Maximum Entropy (ME), and also used successfully the Naive Bayes classification. They examine the classification using the Weka tool and proposed an automated semantic sentiment analysis to detect suicide attempts. Here they perfectly extract the complex sentences also semantic weight of every word, and the content of every tweet data. For that matter of their perform, they explore the text based complex contents which are the tweets related to suicidal attempts Their dataset is built using 892 tweets. And the ultimate result using Naïve Bayes accuracy of 86.09%

Bernert, R.A. et al. [3] evaluated emotional content among suicide attempt notes using natural language processing (NPL). The study utilized supervised knowledge methods, which contained ensemble learning strategies specifically random forests, decision trees, naïve Bayes classification, support vector machines (SVM), and logistic/least squares regression. Used unsupervised learning strategy which included self-organizing maps (SOM), principal component analysis (PCA), clustering algorithms, neural networks, and decision trees. They collected 594 records of data from random user reviews and ultimate result using the decision tree was 89%

Ji, S. et al. proposed a text-based suicide classification to specify whether candidates, through their writings, contain suicidal ideations. Machine learning techniques and NLP have also been applied in this area. they also established a model using SVM, CNN, and ANN. They collected 65,756 data from various social media including (Reddit, Twitter, Facebook, and news portals)

Drew Wilimitis, B.S. et al. [5] Studied an observational crew of adult patients aged greater than or equal to 18 years at VUMC from June 2019 to September 2020, extracted from the Vanderbilt Research Derivative, a clinical study repository. They proposed a VSAIL model that was originally trained to utilize the random forest algorithm on a heterogenous mix of adult VUMC patients

In this research Valeriano, K., Condori-Larico, A. and Sulla-Torres, J. et al. [6] used Text Vector Representation to develop this. This is mainly based on natural language processing techniques and Machine Learning. They managed general statements from tweets. A data set of 2068 text sentences was collected. They're used classification algorithms such as Support Vector Machine (SVM) and Logistic Regression (LR). They also used Word2vec and TF-IDF techniques. The accuracy was 71.4% for TF-IDF and 78.2% for Word2Vec.

Yeskuatov, E., Chua, S.-L. and Foo, L.K. et al. [7] used ML and NLP techniques to detect suicidal ideations. To find out suicidal ideation detection they used a huge amount of

Reddit Data targeting suicidal attempts. In this study, they apply different types of classifiers like TF–IDF, CNN, LDA, LIWC, LR, etc. The ultimate TF–IDF and Word2Vec achieved an accuracy of 84.16%.

Mbarek, A. et al. [8] This study used a natural language processing method based on a machine learning kit for the process of natural language very easily. Which understands the language utilized in a text and reveals the impression behind it. They decided on five classifiers including Adaboost, J48, BayesNet, SMO, and Random Forest. A dataset used Twitter data. The ultimate result was 89% with the classifiers Random Forest and Adaboost.

Barros, J. et al. [9] in this research used a predictive model for suicide risk using data mining (DM) research. They use 777 clinical patient data to analyze targeting suicide and mood disorders. DM and machine-learning tools were used via the support vector machine method. Techniques used in this research are CART, SVM, KNN, CNN, Random Forest, and,AdaBoost.

2.3 Conclusion

All of the analysis was done with the express goal of supporting health and content analyzer and social media manager as well as health management sectors. Identifying and predicting this type of risk is a critical task. Also using this prediction applying in the real life is more challenging.

CHAPTER 3

Theoretical Model

3.1 Introduction

To implement a system for research, a qualitative dataset is needed. In prediction-based analysis, there are some independent factors and a target element. The target factor is dependent on independent elements. Independent factors should be a connection with the target factor. In this section, we will describe the visualization of various factors and also will describe the implementation of the procedure as methodology. Actually, from this chapter, we can know the theoretical background of the following terms which are used in this research.

3.2 Classification Algorithms

The classification algorithms are the supervised understanding method which will be utilized to determine the classification of new statements on the grounds of data which data to be trained. In classification, a program understands the provided dataset and then classifies the new statements into different categories or groups. Classification is the procedure of software understanding a dataset or observations and then classifying new statements into one of multiple categories or classes. In this study, I operated four types of classification algorithms among many machine learning algorithms founded on the previous records of their performance to get satisfactory results. They are the Logistic Regression (LR), and then Decision Tree Classifier, as well as Random Forest (RF), also the Support Vector Machine (SVM) and K-Neighbors Classifiers. The details of them are shown individually below. [18]

3.2.1 Logistic Regression

a statistical method for describing a binary conditional variable in its simplest form using a logistic process. There are many developed versions, though. Regression analysis uses a technique called logistic regression to calculate the parameters of a logistic model (a form of binary regression).

Here they perfectly extract the complex sentences also semantic weight of every word, and the content of every tweet data. For that matter of their performance, they explore the text-based complex contents which are the tweets related to suicidal attempts.

Calculating probabilities utilizing a logistic regression equation is utilized in statistical software to understand clearly and perfectly where is the connection in the dependent variable as well as the additional distinct variables. This condition of analysis can help you in predicting the probabilities of a circumstance or a decision arising.

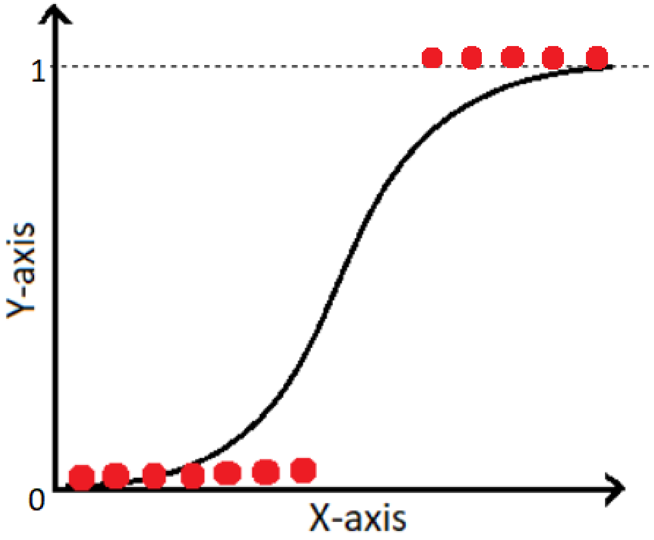


Figure 3.2.1.1: Logistic Regression [19]

An unconditional variable's outcome is indicated by operating logistic regression. And as an outcome, the result must be a discrete or absolute weight. It can be No or Yes, 1 or 0, False or True, and can be so on, but rather than giving detailed values it produces probability values. Instead of fitting a regression line, we suit a logistic operation in logistic regression, which predicts the two highest weights like (0 or 1). Finally, we can tell that this is a classification method that leverages the concept of predictive modeling as regression.

3.2.2 Decision Tree Classifier

The decision tree classifier is a class differentiator that splits the activity set recursively until each section contains only or primarily

samples from one class. The decision Tree algorithm belongs to the relative of supervised understanding algorithms. Like the different supervised learning algorithms, the decision tree algorithm can be utilized for solving regression and classification issues too. The objective of using a Decision Tree is to develop a training model that can be utilized to indicate the class or value of the mark variable by understanding uncomplicated determination rules inferred from initial data. In decision trees, for pointing a class label for a record we start from the root of the tree. Then we resemble the values of the root detail with the record's feature. After based on the comparison, we follow the branch corresponding to that weight and bounce to the next following node.

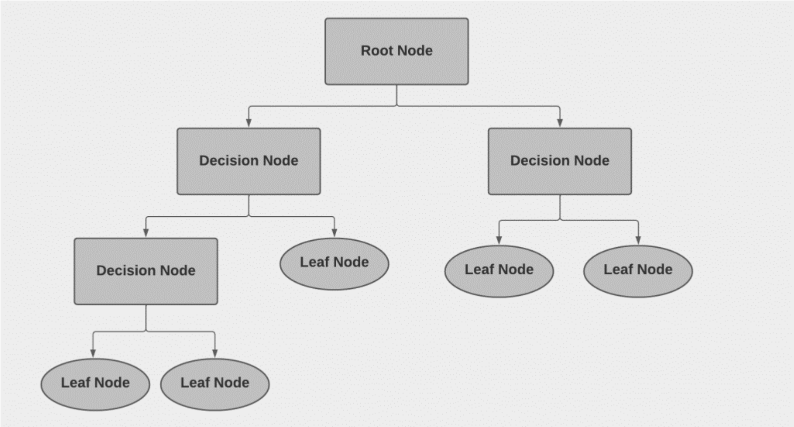


Figure 3.2.2.1: Decision Tree Classifier [21]

In the flowchart, the design of the internal nodes shows the difficulties or properties at a particular class or level. Every branch denotes a particular output, whereas the path from the leaf to the root defines categorization criteria. The most significant part of the learning algorithm based on numerous learning approaches is decision trees. They have improved the accuracy, stability, and readability of prediction models.

3.2.3 Concept of Random Forest (RF)

Random forest (RF) is a powerful machine learning method developed by Leo Breiman and Adele Cutler that combines the outcome of numerous decision trees for creating a single outcome. Random forest is mainly a supervised easy-going technique and learning algorithm that is utilized for both classifications as well as regression. Random Forest is more effective and better than a single decision tree. The hidden reason is it reduces the over-fitting by averaging the result. We will able to understand the functional stages of the Random Forest algorithm also the assistance of the subsequent stages given below:

First, let's begin by selecting randomly from a pre-provided dataset. Then, this RF algorithm will construct a decision tree for every sample. Then it will obtain the projection output from every decision tree. After that, voting will be conducted for every expected outcome. Lastly, it selects the tallest-voted projection outcome as the ultimate projection outcome.

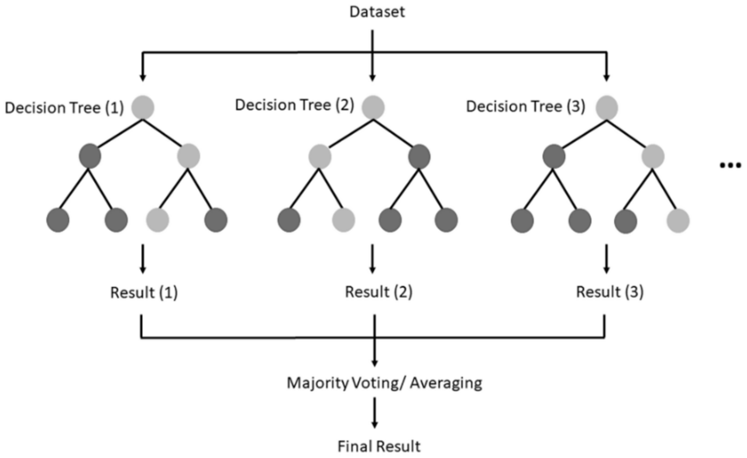


Figure 3.2.3.1: Random Forest (RF) [22]

The (RF)Random Forest algorithms are formed of a set of decision trees, and each tree in the ensemble is made up of a bootstrap model, which is a data model obtained from a training set with replacement. Almost One-third of the training selected sample is placed aside as test data, directed to the out-of-bag instance, which we'll consult later. Utilizing attribute bagging, the other instance of the random values needs to be injected into the dataset, enhancing the dataset's variety and reducing the correlation between decision trees. The forecast will be specified differently depending on the type of hardship.

3.2.4 Support Vector Machine (SVM)

The support vector machine is a supervised-based machine learning algorithm. Which can be utilized for meeting both classification and regression difficulties correctly. This is especially utilized in decoding classification difficulties. Through this support vector machine algorithm, we can plot each data item as a point in an n-dimensional margin by using the weight of every element existing with certain lines. After That, we conduct the category by discovering the correct hyperplane that distinguishes the 2 styles of categories nicely.

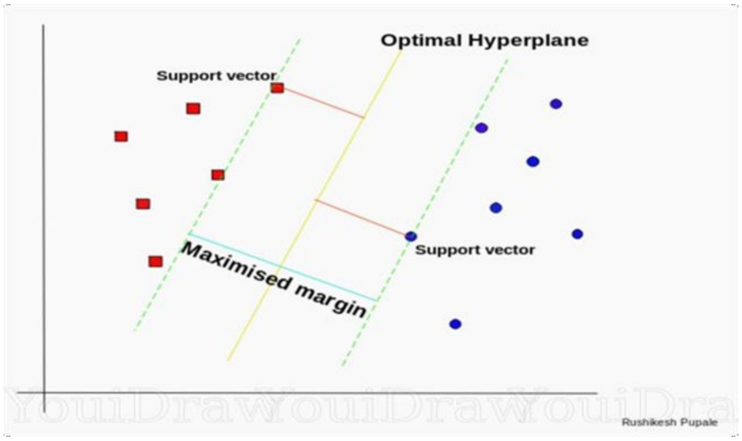


Figure 3.2.4.1: support vector machine (SVM) [20]

3.2.5 K-Nearest Neighbor (KNN)

One of the simplest machine learning algorithms, based on the supervised learning methodology, is K-Nearest Neighbor. The K-NN algorithm takes into account how similar the new case/data is to unconstrained cases and classifies the most recent case into those categories. The k-NN technique provides all the data that is readily available and categorizes a new data point using the parallel. This shows that when new data is present, it can typically be classified using the K- NN algorithm into a suitable class. Although the K-NN technique can be used for both classification and regression, its primary application is for classification problems. Since K-NN is a non-parametric technique, it makes no assumptions about the underlying data.

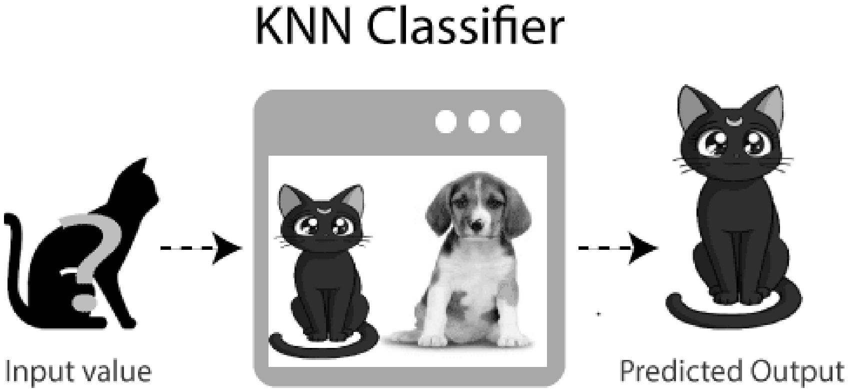


Figure 3.2.5.1: K-Nearest Neighbor (KNN) [24]

3.2.6 Concept of deep learning

Deep learning is called a subset of machine learning, which is a neural network with three or sometimes more extra layers. These neural networks attempt to affect or affect the conduct of the mortal brain albeit distant from reaching its capacity to allow it to “retain” enormous quantities of data. Deep learning doesn't indicate the machine learns more in-depth understanding. It means a machine utilizes various layers to understand the data. In deep learning, the learning step is done via a neural network. Where a neural network is an architecture where the layers are piled on top of each other. Most of the deep learning techniques utilize neural network architectures. For that reason, most deep learning pinnacles are usually directed to as deep neural networks. The phrase “deep” mainly directs to the number of hidden layers in the neural network. Standard neural networks just hold two to three secret layers, whereas deep networks can have as numerous as 150. Deep learning instances are prepared by utilizing extensive collections or groups of marked data and neural network architectures that understand elements straight from the data without the requirement for manual component extraction.

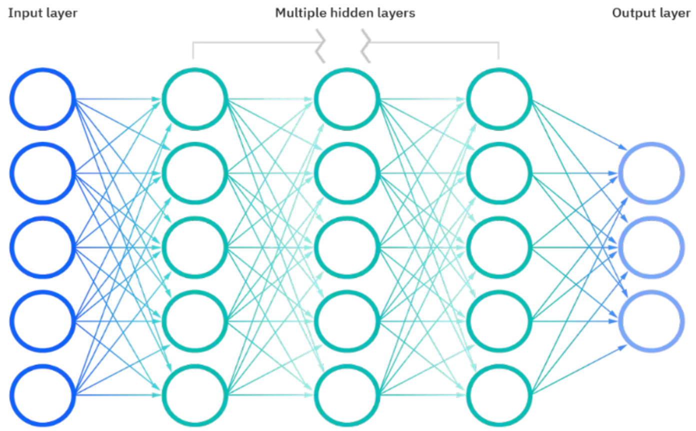


Figure 3.2.6.1: Deep learning matrix [23]

3.3 Conclusion

In the above, all of the theoretical facts about this study in this chapter are attempted to explain. After That, we conduct the category by discovering the correct hyperplane that distinguishes the 2 styles of categories nicely. Also, this chapter made a huge discussion regarding various effective latest used classification algorithms in them one of the effective classifiers is Decision Tree Classifiers, Random Forest, other hands K-Nearest Neighbor & Logistic Regression in detail. Hopefully, there does no doubt about the above facts.

CHAPTER 4

Experimental Model

4.1 Introduction

In any type of research, the experimental result is very essential. All the researchers want to reach the highest accuracy level according to their work. This accuracy level may be the difference by using various algorithms and methodologies. The researchers specify the algorithm and methodology which provide the most beneficial accuracy level for the related research.

4.2 Proposed Model

For any type of research, a structured dataset is needed to understand. In this research, we have collected a dataset from Kaggle. For using this research, we have to process the text and then convert it to numerical value across the individual text.

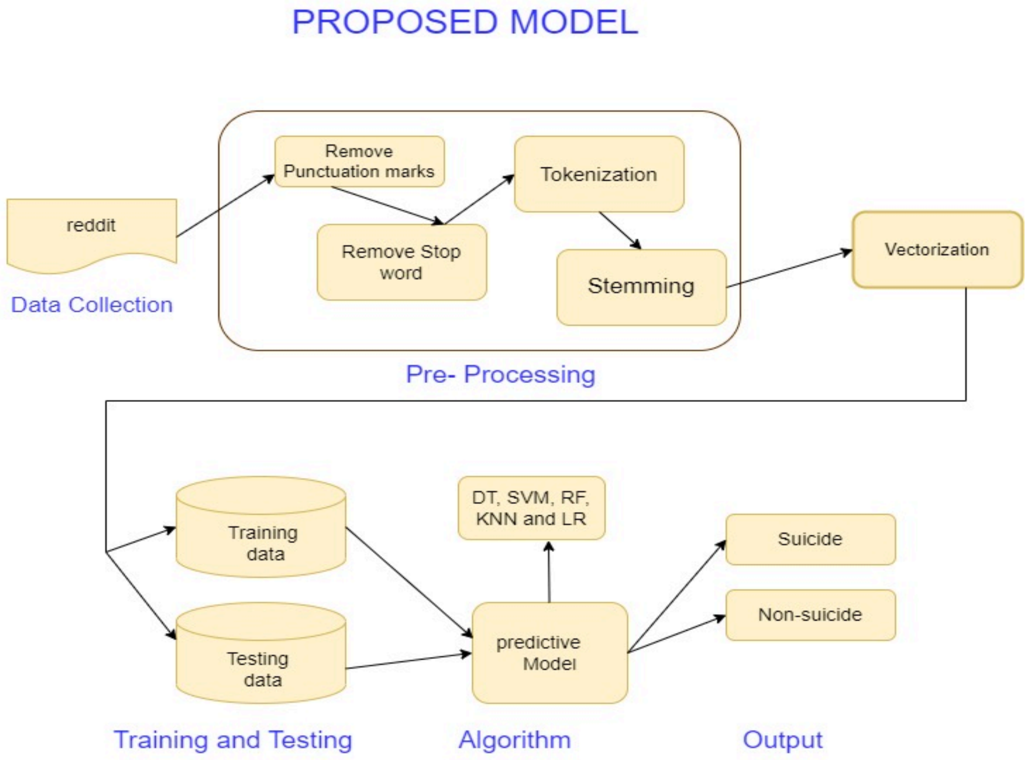
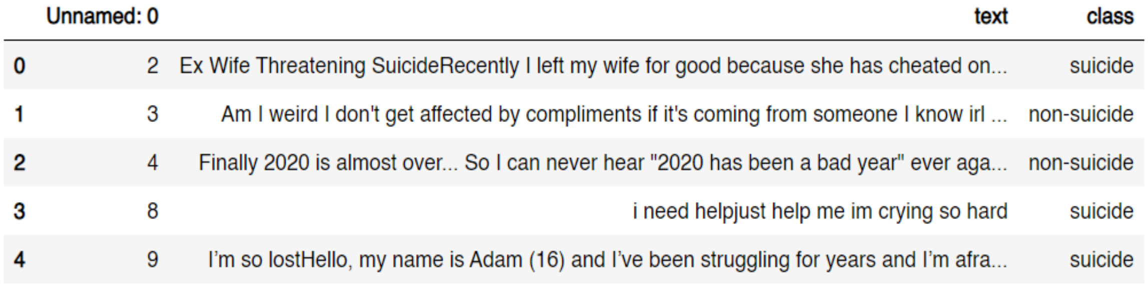


Figure 4.2.1: Proposed model

The summary of the procedure illustrated in this model also describes the overall idea of how effectively and correctly this prediction will ensue. Firstly, we gathered data from an online source called Kaggle where various kinds of authentic data are preserved. After that, we develop a supervised machine learning which prepares the data and extracts information from the data. In our experimental case, we predict Suicide and Depression Detection. We furthermore use machine learning methods to train the learning. In this method, we predict climate change and then we discover the accuracy ratio. Each part of our proposed method is illustrated in the following section.

4.3 Dataset

As a dataset, here utilized the Suicide and Depression Detection dataset that can be used to detect suicide and depression in a text from Kaggle containing 2,33,338 user records. The dataset is a collection of posts from the "SuicideWatch" and "depression" subreddits of the Reddit platform. The posts are collected utilizing Push shift API. All posts that were made to "SuicideWatch" from Dec 16, 2008(creation) till Jan 2, 2021, were collected while "depression" posts were managed or collected from Jan 1, 2009, to Jan 2, 2021.

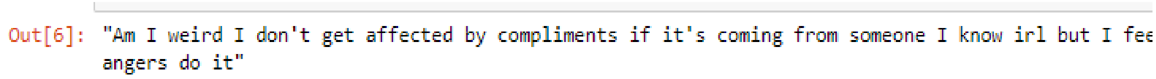


	Unnamed: 0	text	class
0	2	Ex Wife Threatening SuicideRecently I left my wife for good because she has cheated on...	suicide
1	3	Am I weird I don't get affected by compliments if it's coming from someone I know irl ...	non-suicide
2	4	Finally 2020 is almost over... So I can never hear "2020 has been a bad year" ever aga...	non-suicide
3	8	i need helpjust help me im crying so hard	suicide
4	9	I'm so lostHello, my name is Adam (16) and I've been struggling for years and I'm afra...	suicide

Figure 4.3.1 Dataset sample

4.4 Data Transformation

Data transformation or Pre-processing is a method that is used to convert the raw data into a useful and efficient format before feeding it to the algorithm. To assure the quality of data, it can directly affect the ability of our model to learn if not processed. Sometimes the reviews may include extreme and insignificant data for research. And hence they require some processing. The following example represents a snippet of a text word that consists of text along with punctuations, stop words, etc. we can take examples from our data Such as:



```
Out[6]: "Am I weird I don't get affected by compliments if it's coming from someone I know irl but I fee
        angers do it"
```

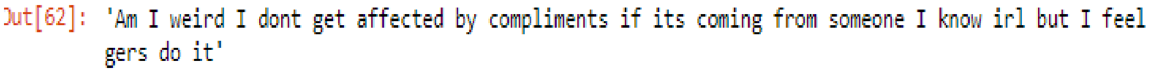
Figure 4.4.1 Sample data pre-processing

4.4.1 Remove Outliers from Dataset

There have been seen many awkward and extra values for different attributes. It seems un realistic to me. For gaining more satisfactory performance for the accurateness rate, those values are fired from different attributes of this dataset and assembled the dataset sensible to all by its containing values. Dropped the column called “Unnamed 0:” because it hasn't any practical role in this research

4.4.2 Punctuation Character Removal

The English language utilizes many punctuation characters in texts which carries a little matter in sentiment polarity. Punctuation varies to get the correct impression from the text. So, a simple script was used to strip all punctuation characters clear from the data saw the flowing example:



```
Out[62]: 'Am I weird I dont get affected by compliments if its coming from someone I know irl but I feel
         gers do it'
```

Figure 4.4.2.1 Punctuation Character Removal

4.4.3 Stop Words Removal

A stop word is a word that seems usually in the dataset despite having no sentiment contradiction associated with it. This can be purified before or after the processing of natural language data. In sentiment analysis, many of the words in English like

'about', 'above', 'after',

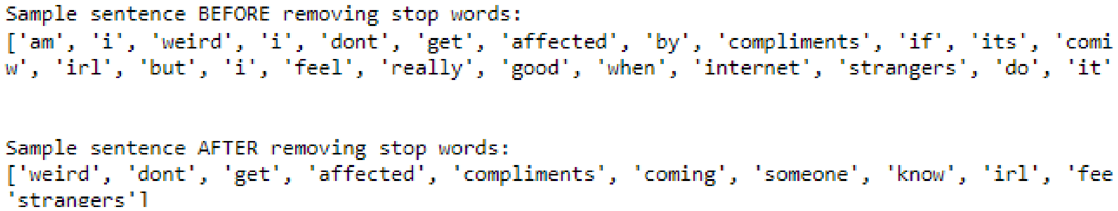
'again',

'against',

'ain',

'all',

etc. least significance or do not convey any meaning. Which can discard from the input text seen in the flowing example. As the overall polarity of a study does not depend on those words.



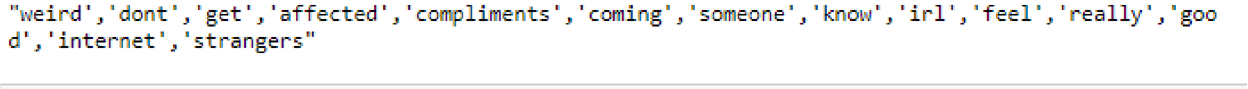
```
Sample sentence BEFORE removing stop words:
['am', 'i', 'weird', 'i', 'dont', 'get', 'affected', 'by', 'compliments', 'if', 'its', 'comi
w', 'irl', 'but', 'i', 'feel', 'really', 'good', 'when', 'internet', 'strangers', 'do', 'it'


Sample sentence AFTER removing stop words:
['weird', 'dont', 'get', 'affected', 'compliments', 'coming', 'someone', 'know', 'irl', 'fee
'strangers']
```

Figure 4.4.3.1 data pre-processing

4.4.4 Tokenization

Tokenization directs to a procedure by which a portion of sensitive data is. Such as a David card number, which is replaced by a surrogate weight understood as a token. Text segmentation or Tokenization is the method of separating the composed text into meaningful units, such as words, sentences, or topics. As an example, after applying Stop Words Removal and Tokenization to our dataset we see:



```
"weird','dont','get','affected','compliments','coming','someone','know','irl','feel','really','goo
d','internet','strangers"
```

Figure 4.4.4.1 Tokenization

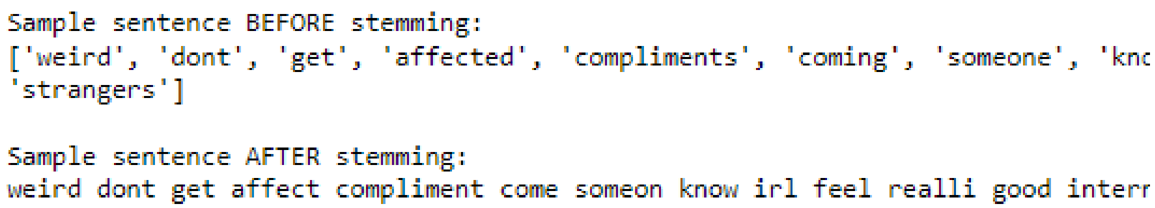
4.4.5 Steaming

Stemming is a standard data transformation process operation for Natural Language Processing (NLP) tasks. Stemming programs are generally directed to as stemming algorithms or stemmers.

There are two types of steaming algorithm are there:

1. Porter stemming algorithm
2. Lancaster stemming algorithm

After using porter steaming algorithm



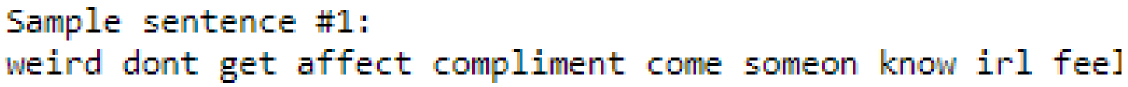
```
Sample sentence BEFORE stemming:
['weird', 'dont', 'get', 'affected', 'compliments', 'coming', 'someone', 'knc
'strangers']

Sample sentence AFTER stemming:
weird dont get affect compliment come someon know irl feel realli good interr
```

Figure 4.4.5.1 Steaming

4.4.6 Vectorization

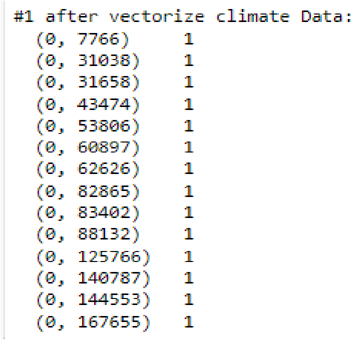
In Machine Learning, vectorization is a term in element extraction. The concept is to obtain some distinct features out of the text for the model to train on, by transforming or converting text to numerical vectors. Let's look into 1 sample sentence to understand better what vectorization does



```
Sample sentence #1:
weird dont get affect compliment come someon know irl feel
```

Figure 4.4.6.1 Numerical vector

this sentence has a couple of words in common - "come", "someon"



```
#1 after vectorize climate Data:
  (0, 7766)	1
  (0, 31038)	1
  (0, 31658)	1
  (0, 43474)	1
  (0, 53806)	1
  (0, 60897)	1
  (0, 62626)	1
  (0, 82865)	1
  (0, 83402)	1
  (0, 88132)	1
  (0, 125766)	1
  (0, 140787)	1
  (0, 144553)	1
  (0, 167655)	1
```

Figure 4.4.6.1: Vectorization

Based on the column size of our vectorized data, we can notice there were 177561 tweets in the dataset and 67195 individual unique words.

4.4.7 Bi-Grams

Using N-Grams, we can group N numbers of words and investigate their frequencies for exact sentiment ratings.

- Top 10 Occurrences of Bi-Grams of “**suicide**” detection Tweets

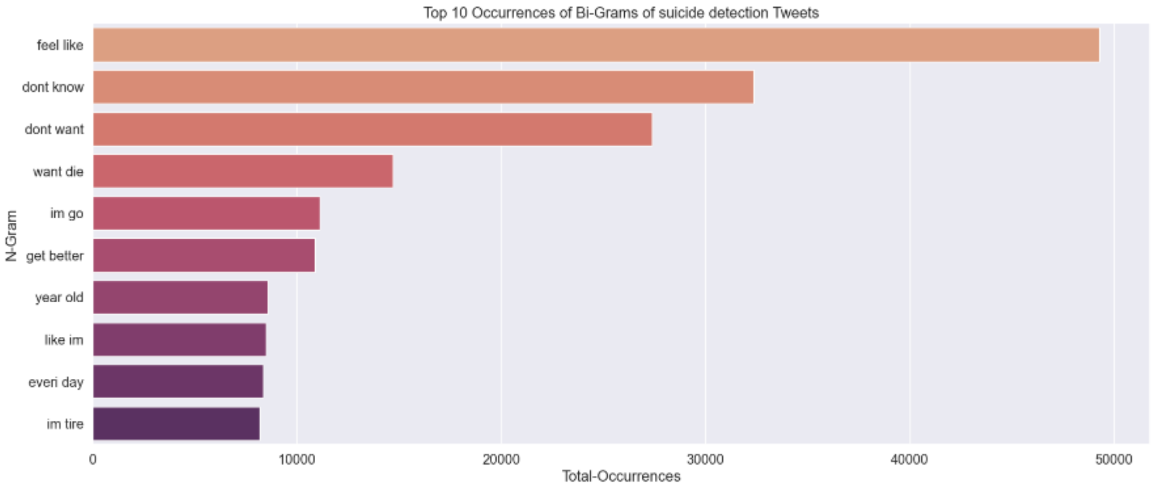


Figure 4.4.7.1: N-Gram (Opinion: suicide)

- Top 10 Occurrences of Bi-Grams of “**non-suicide**” detection Tweets

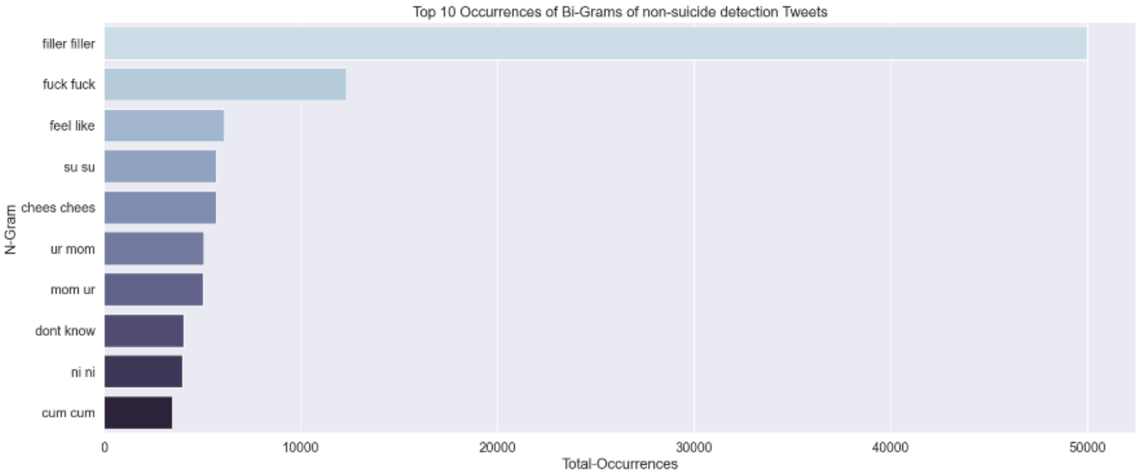


Figure 4.4.7.2: N-Gram (Opinion: non-suicide)

4.4.8 Tri-Grams

Let's try tri-grams and see if it finds more meaningful combinations of words than bi-grams**.**

- Top 10 Occurrences of Tri-Grams of “**suicide**” detection Tweets

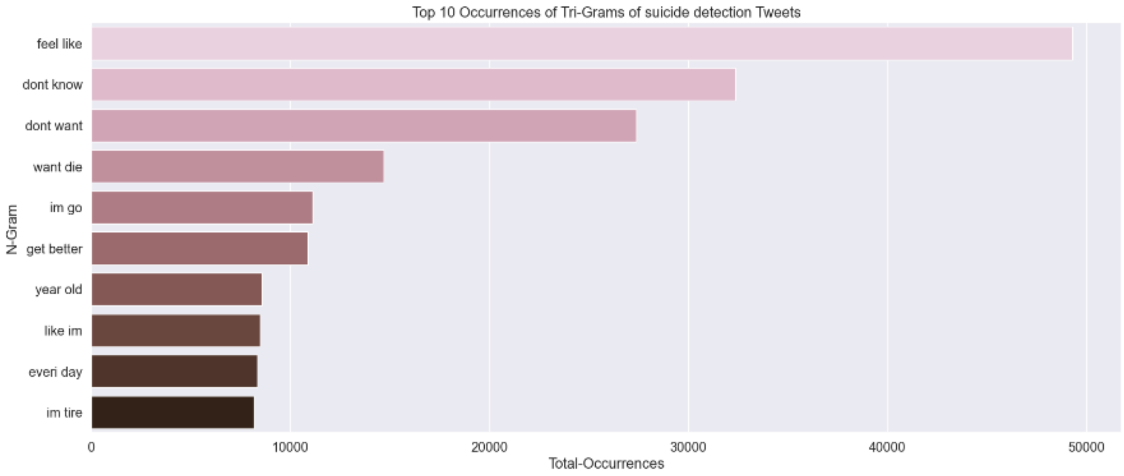


Figure 4.4.8.1 Tri-gram-Gram (Opinion: suicide)

- Top 10 Occurrences of Bi-Grams of “**non-suicide**” detection Tweets

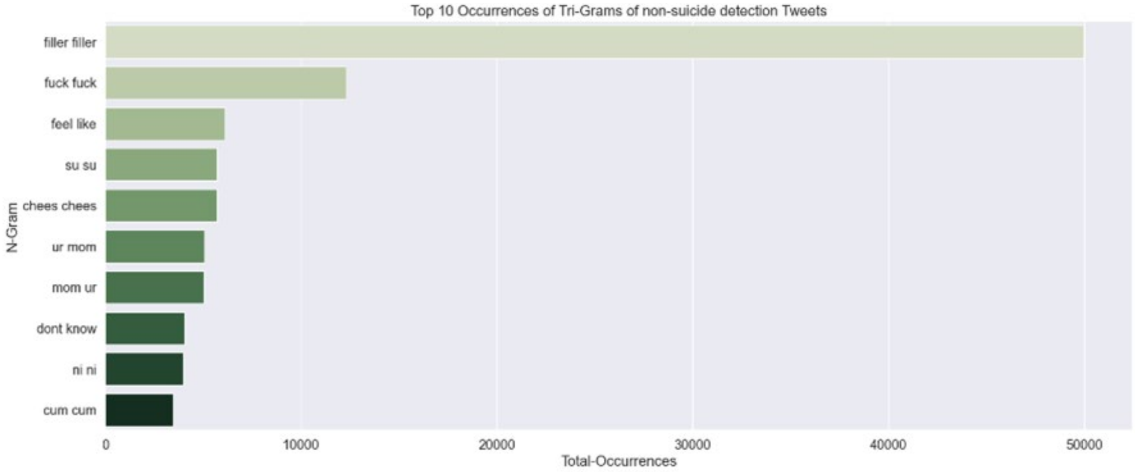


Figure 4.4.8.2 Tri-gram-Gram (Opinion: non-suicide)

4.5 Conclusion

In this chapter, utilizing data mining processes are discussed like data simplification, modification, visualization, and many more. The most analytical part of this chapter is the processing using bi-gram tri-gram.

CHAPTER 5
Result & Discussion

5.1 Introduction

In this chapter, the process of finding out the ultimate result of this study will be discussed. For that cause, some measures (preprocessing & declaration) need to debate for using ML algorithms. After applying ML algorithms, the implementation measured parameters will be analyzed like accuracy, recall, F-Score, confusion matrix, ROC curve, and many more for estimating the best outcome of this study.

5.2 Environment

We conducted the experiments on a computer with 4GB of RAM and 5 Intel CPUs (2.43GHz each). Anaconda Navigator was used in our studies to assess the suggested categorization models and comparisons. We used 4 algorithms to understand the dataset. These include DT, LR, RF, and RF. Using several algorithms, we obtained a variety of results.

5.3 Applying Machine Learning Algorithms

In order to get a more reasonable accuracy rate for the classification, five ML methods are implemented based on the performance of past research for categorizing the target class (Suicide). Logistic Regression (LR), Decision Tree Classifier (DT), Support Vector Machine (SVM), Random Forest (RF), and K-Nearest Neighbors Classifier are the classification algorithms (KNN).

5.3.1 Analysis of Performance Measured Used

Here, to apply classification ML algorithms, percentage division is used as a data mining method. The percentage split is a resampling method in which n% of the rows are set aside as the training dataset for building the model and (n-100) % of the rows are set aside as the trial dataset for testing the model. In contrast to the learned data, the target classifier is trained. The classification accuracy, on the other hand, is assessed using the trial dataset.

In this research we set the percentage split, 80% of the rows are utilized as the training dataset for constructing the model, and the remaining 20% as the trial dataset for testing the model purpose.

5.3.2 Data Visualization

Data visualization is an essential preprocessing job, which operated a graphical model to simplify and understand easily the overall status of complex data. Visualization methods have been newly used to visualize online learning factors. Instructors can use graphical presentations to understand their learners nicely and become conscious of what is happening in distance lessons. This study visualizes the existing data set utilizing the Anaconda Navigator tool. Actually, we spilled data 3 times for memory shortage. As illustrated in Figure 4.1, the data set is pictured based on sentiment into non-suicide risk 38713 and suicide risk 38645 To understand the dataset for machine learning objectives, we have left suicide as 0, non-suicide as 1. Here, 80 % data was utilized for training, and 20% of the data was used for testing for every model this work

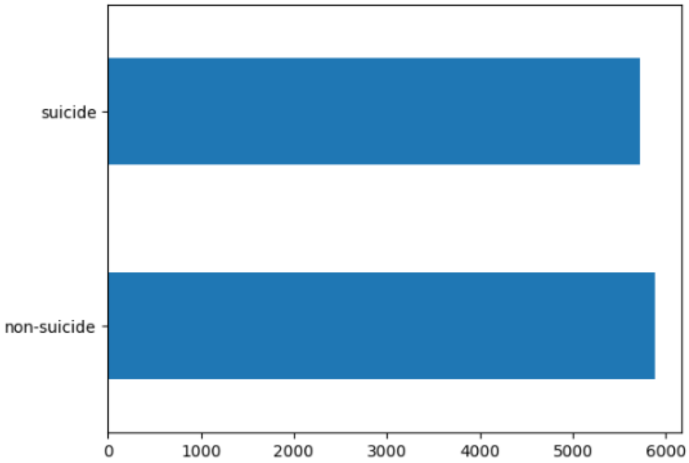


Figure 5.3.2.1 Data visualization

5.3 Analysis of Performance Measured Used

Used For measuring the performance applying to classification algorithms (Logistic Regression, Decision Tree Classifier, Random Forest, Support Vector Classifier, And K-Neighbors Classifier) of the declared model, performance-measured parameters (accuracy, precision, recall, F-Score, confusion matrix) are used. The results for these performance-measured parameters are displayed below.

5.3.2 Accuracy Rate of Classification

Here in this study, the most significant performance calculated parameter is the accuracy rate. Truthfully, it is estimated by splitting the number of precisely categorized sample models by the entire number of samples multiplied by 100. The sum of True-Positive (TP) and True-Negative (TN) is the exact classified sample

$$\mathbf{Accuracy} = \frac{\mathbf{TP+TN}}{\mathbf{TP+TN+FP+FN}} \times \mathbf{100} \qquad (1)$$

The obtained accuracy rates of these classification algorithms are given below:

Table 5.3.2.1: Accuracy rate of classification

Applied ML Algorithms	Accuracy Rate
Logistic Regression (LR)	92.87%
Random Forest (RF)	88.57%
Decision Tree Classifier (DT)	84.50%
Support Vector Machine (SVM)	92.12%
K-Nearest Neighbors (KNN)	73.69%

According to the table 5.3.2.1, it can declare that Logistic Regression Classifier got the highest accuracy rate for this proposed model.

5.3.3 Precision

Precision is a performance metric utilized for pattern recognition and classification processes in machine learning. Precision is one of the measurements of a machine learning model's interpretation. The accuracy of a model's positive prediction. Precision is defined as the percentage of true-positive examples to predicted yes data, according to the Confusion Matrix

$$\textbf{Precision} = \frac{\textbf{TP}}{\textbf{TP} + \textbf{FP}} \qquad (2)$$

The obtained accuracy rates of these classification algorithms are given below:

Table 5.3.3.1: Precession rate of classification

Applied ML Algorithms	Precision Rate
Logistic Regression (LR)	92.94%
Random Forest (RF)	88.57%
Decision Tree Classifier (DT)	84.50%
Support Vector Machine (SVM)	92.14%
K-Nearest Neighbors (KNN)	92.12%

5.3.4 Recall

Recall is a parameter that considers how nicely a standard can detect positive models. Recall is another name for sensitiveness. Recall is described as the ratio of true-positive models to genuine yes samples, according to the Confusion Matrix

$$\textbf{Recall} = \frac{\textbf{TP}}{\textbf{TP} + \textbf{FN}} \qquad (3)$$

The obtained accuracy rates of these classification algorithms are given below:

Table 5.3.4.1: Recall rate of classification

Applied ML Algorithms	Recall Rate
Logistic Regression (LR)	92.87%
Random Forest (RF)	88.57%
Decision Tree Classifier (DT)	84.50%
Support Vector Machine (SVM)	92.12%
K-Nearest Neighbors (KNN)	73.69%

5.3.5 F-Score

The F-Score is also understood as the F1-Score or the F-Measure. Using both recall and accuracy, the F-Score can supply a better useful measure of test performance. When the F-Score score reaches 1, it means that both recall and accuracy are ideal

$$\textbf{F1 score} = (2 * \textbf{Precision} * \textbf{Recall}) / (\textbf{Precision} + \textbf{Recall}) \tag{4}$$

This section, we will explain the results that we obtained of our applied techniques.

The obtained accuracy rates of these classification algorithms are given below:

Table 5.3.5.1: F1 Score of classification

Applied ML Algorithms	F1 Score
Logistic Regression (LR)	92.86%
Random Forest (RF)	88.57%
Decision Tree Classifier (DT)	84.50%
Support Vector Machine (SVM)	92.12%
K-Nearest Neighbors (KNN)	72.56%

5.5 Confusion Matrix

The Confusion Matrix is a tool for displaying a model's performance or how a model generated a prediction in Machine Learning. The Confusion Matrix allows us to see where our model becomes confused while deciding between two classes. A 2×2 matrix may be used to visualize it, with the row representing the actual truth labels and the column representing the prediction labels

Table 5.5.1: Basic Diagram of a Confusion Matrix

	Positive	Negative
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)

Here, the Confusion Matrix is plotted below for all four ML algorithms. Hope, it will easy to understand after viewing the figure 5.5.1.

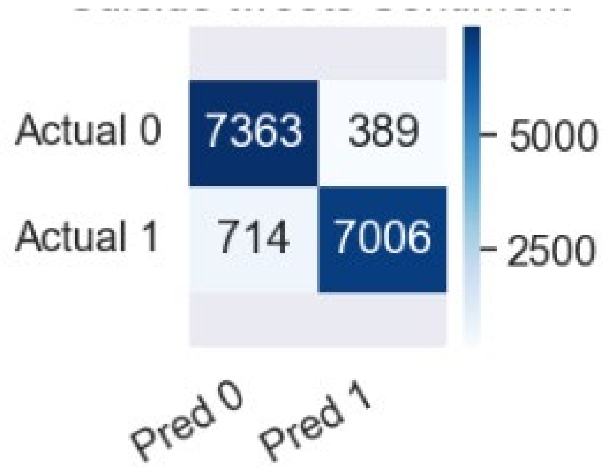


Figure 5.3.5.1: Confusion matrix of Logistic Regression

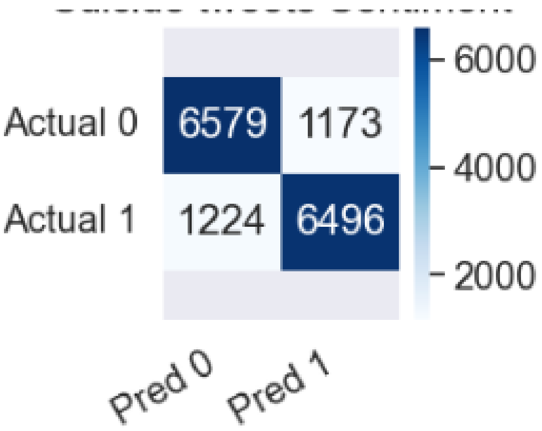


Figure 5.5.2: Confusion matrix of Decision Tree Classifier

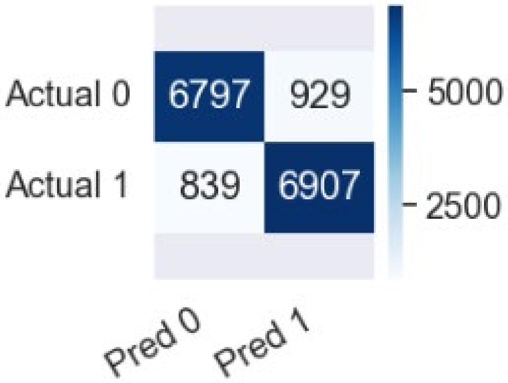


Figure 5.5.3: Confusion matrix of Random Forest Classifier

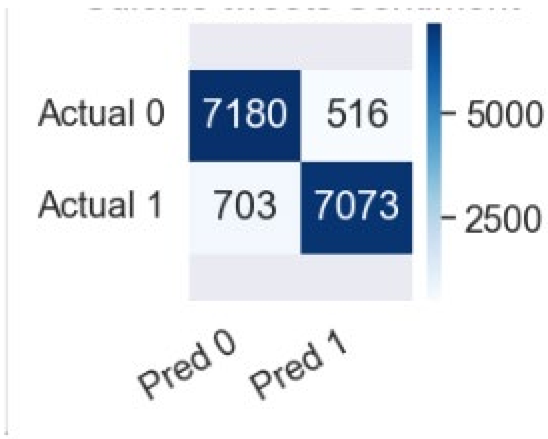


Figure 5.5.4: Confusion matrix of Support Vector Machine

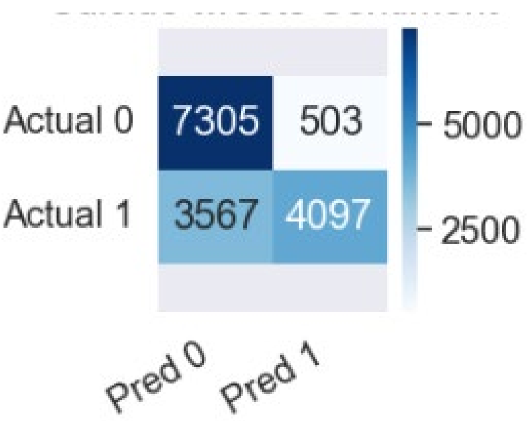


Figure 5.5.5: Confusion matrix of K-Neighbor Classifier

5.6 Final Result and analysis chart

In this figure below accuracy and precision, recall and f1-score results of the best four machine learning techniques (DT, SVM, RF, KNN and LR) is given below:

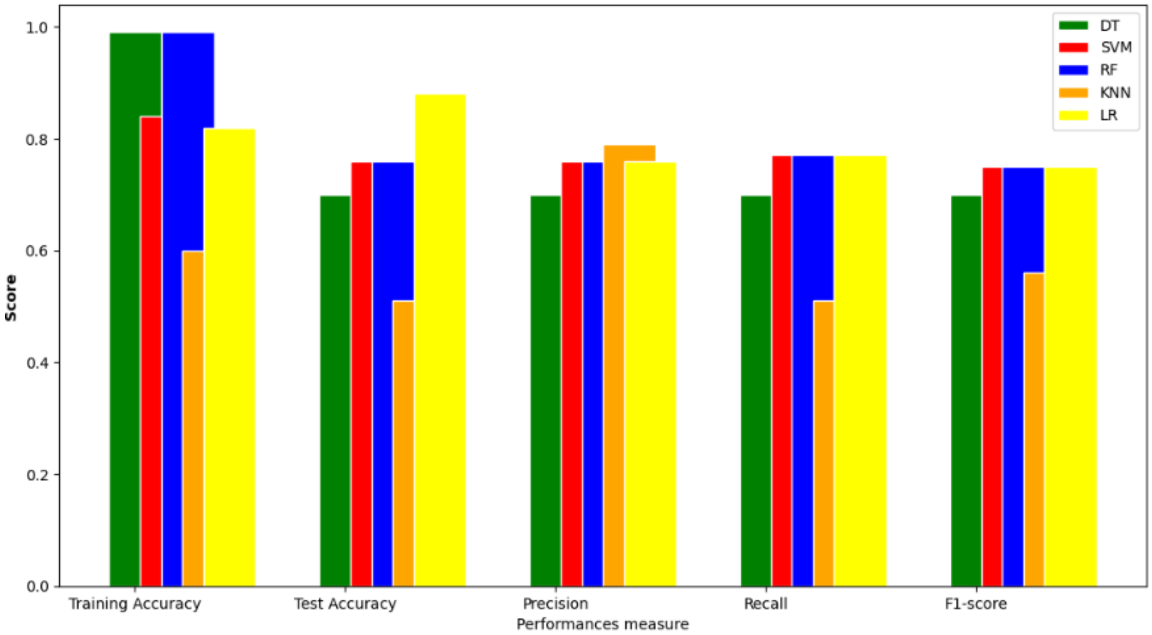


Figure-5.6.1: Comparison of Accuracy Result among DT, SVM, RF, KNN and LR

Using Long short-term memory is training and test accuracy is given below:

Table-5.6.2: Comparison of Accuracy among DT, SVM, RF, KNN and LR

Measure	DT	SVM	RF	KNN	LR
Training Accuracy	99.54%	84.22%	99.54%	60.08%	82.88%
Test Accuracy	70.47%	76.97%	76.56%	51.00%	76.84%

In Figure-5.7 and Table-5.3 is the comparison of accuracy result among DT, SVM, RF, KNN and LR. Where Logistic Regression scored higher.

5.6 Limitations

In our investigation, we employed only 232074 reddit post. This works is effective, but since just a small dataset is used, accuracy and F1 score are insufficient. Large data sets are needed for logistic regression (LR), which is extremely expensive to train. Using computers fitted with pricey GPUs, these sophisticated models may be trained over the course of weeks. When working with short sentences, using Logistic Regression (LR) for sentence classification is an intriguing strategy, but as sentences get longer, recurrent networks should be a more suitable technique.

5.6 Conclusion

In this chapter, it is marked that the Support Vector Machine Classifier & Logistic Regression (LR) both executed fantastically in almost all sectors (Accuracy, Precision, Recall, F-Score, Confusion Matrix & final result curve). But, in the measure of precision rate, Logistic Regression (LR) functioned well than Decision Tree Classifier and acquired the tallest accuracy rate of 92.87% on this clustering dataset which is included in the experimental model chapter. Overall, the performance & result of this model fulfilled the requirements of this study.

CHAPTER 6
Critical Appraisal

6.1 Introduction

In this chapter, the strength, languish, and content of my study will discuss broadly. And this will facilitate the other experimenters to evaluate this study or research purpose. Other researchers will be able to utilize its resource very efficiently

6.2 SWOT Analysis

SWOT analysis is a strategic practice and strategic management technique that may be utilized to help an individual or institution discover its analysis planning powers, drawbacks, possibilities, and dangers. It's also understood as system analysis or situational evaluation.

6.2.1 Strength

In this study, the primary strong point is the goal of this research outcome. Honestly. this analysis is founded on suicide detection based on the concern for human beings. The suggested predictive measure benefits overcome this issue by detecting early suicide issues where people who are at risk can be rescued with the help of machine learning. This study also benefits social media engineers to use machine learning to predict the risk of suicide. This research also builds a connection between the health sector and the Data Mining field focused on early suicide detection where the earth is pushing to rely on computer science day by day. In my opinion, it is the strength of my research work.

6.2.2 Drawback

If I address the deficiency of this study work, then it will be the type of data that has been used finally after categorizing it as a dataset. The ultimately utilized dataset for creating this model is categorical. This instance can be executed when the input data would be separated into many categories for the features utilized in the dataset. In my point of view, it may be a drawback. But the most remarkable point is that as data mining researchers, on today's planet people can categorize any data like numerical or, distinct data. So, it cannot be a primary issue behind this study.

6.2.3 Opportunity

This study has tremendous possibilities and opportunities. As today every people connected to social media, and they express their impressions. With this predictive model, social media engineers can use machine learning to predict the risk of suicide or depression. Also utilizing this predictive model, doctors or, diagnostic centers can predict the risk and cause of suicide, and the analyzer will be able to find out the exact number of people who are depressed and going to suicide. In my point of view, it will be a wonderful option for both medical science and data mining researchers as well as social media engineers.

6.2.4 Threat

In my point of view, the most challenging threat is accurately detecting and predicting suicide risks. Because according to this model the highest accuracy rate of suicide detection is 92.86%. So, there has a 08.14 % probability of delivering the wrong detection to people. But it can be enriched or improved by further research shortly.

6.3 Conclusion

After the SWOT analysis of my study work, I can confidently state that this developed predictive sample can play a tremendous role in both the medical sector and as well as the social commercial sector and for the social site manager.

CHAPTER 7
Conclusion

7.1 Conclusion

In this analysis, an innovative and creative ensemble approach is shown by integrating data mining methods. Evaluating classification approaches such as Logistic Regression (LR), Decision Tree Classifier (DT), Random Forest (RF), and Support Vector Machine (SVM) to define which is the most useful at accurately predicting early suicide risk. The absolute outcome illustrates that Logistic Regression that acquired the most increased accuracy rate of 92.86% founded on this proposed predictive model.

The accuracy, precision, recall, F-Score, and confusion matrix are among the six metrics utilized to evaluate the proposed model. The presented model's interpretation was compared to that of different existing models. Based on the results of the experimentation, we can finish that the proposed method improves suicide risk accuracy. The proposed model was designed using Jupyter Notebook (a Python-based IDE) and trained to utilize Kaggle's standard Suicide and Depression Detection, which has 2,32,074 unique records.

In today's society, defining the reason for a problem like suicide is critical. Early suicide detection can help to prevent premature death. The suggested prediction model handles this issue by specifying or identifying suicide risk at an early stage, entitling those who are at risk to an initial step that can be taken to save their life.

7.2 Further Suggested Work

In the future, analysis can be performed to enhance prediction accurateness by integrating various algorithms. Besides, it will be able to concentrating on enhancing classification accurateness and will need to find out the most useful data mining technique utilizing several machine learning algorithms. To do so, our data set may be experimented in a variety of ways. For further research, more well-known suicide datasets may be selected to use. It also decided that it would like to utilize our research methods not just in the field of Suicide, but also in the field of other sentiment.

REFERENCES

[1]T. Sravanthi, V. Hema, S. Tharun Reddy, K. Mahender, and S. Venkateshwarlu, “Detection of Mentally Distressed Social Media Profiles Using Machine Learning Techniques,” *IOP Conference Series: Materials Science and Engineering*, vol. 981, p. 022056, Dec. 2020, doi: 10.1088/1757-899x/981/2/022056.

[2]M. Birjali, A. Beni-Hssane, and M. Erritali, “Machine Learning and Semantic Sentiment Analysis based Algorithms for Suicide Sentiment Prediction in Social Networks,” *Procedia Computer Science*, vol. 113, pp. 65–72, 2017, doi: 10.1016/j.procs.2017.08.290.

[3]R. A. Bernert, A. M. Hilberg, R. Melia, J. P. Kim, N. H. Shah, and F. Abnousi, “Artificial Intelligence and Suicide Prevention: A Systematic Review of Machine Learning Investigations,” *International Journal of Environmental Research and Public Health*, vol. 17, no. 16, p. 5929, Aug. 2020, doi: 10.3390/ijerph17165929.

[4]S. Ji, S. Pan, X. Li, E. Cambria, G. Long, and Z. Huang, “Suicidal Ideation Detection: A Review of Machine Learning Methods and Applications,” *arXiv:1910.12611 [cs]*, Sep. 2020, doi: 10.1109/TCSS.2020.302146.

[5]D. Wilimitis *et al.*, “Integration of Face-to-Face Screening With Real-time Machine Learning to Predict Risk of Suicide Among Adults,” *JAMA Network Open*, vol. 5, no. 5, pp. e2212095–e2212095, May 2022, doi: 10.1001/jamanetworkopen.2022.12095.

[6]K. Y. Valeriano Valdez, J. Sulla-Torres, and A. Condori-Larico, “Detection of Suicidal Intent in Spanish Language Social Networks using Machine Learning,” *researchgate*, Jan. 2020, doi: 10.14569/IJACSA.2020.0110489.

[7]E. Yeskuatov, S.-L. Chua, and L. K. Foo, “Leveraging Reddit for Suicidal Ideation Detection: A Review of Machine Learning and Natural Language Processing Techniques,” *International Journal of Environmental Research and Public Health*, vol. 19, no. 16, p. 10347, Aug. 2022, doi: 10.3390/ijerph191610347.

[8]A. Mbarek, S. Jamoussi, A. Charfi, and A. Ben Hamadou, “Suicidal Profiles Detection in Twitter,” *Proceedings of the 15th International Conference on Web Information Systems and Technologies*, 2019, doi: 10.5220/0008167602890296.

[9]J. Barros *et al.*, “Suicide detection in Chile: proposing a predictive model for suicide risk in a clinical sample of patients with mood disorders,” *Revista Brasileira de Psiquiatria*, vol. 39, no. 1, pp. 1–11, Oct. 2016, doi: 10.1590/1516-4446-2015-1877.

[10]N. J. Carson *et al.*, “Identification of suicidal behavior among psychiatrically hospitalized adolescents using natural language processing and machine learning of electronic health records,” *PLOS ONE*, vol. 14, no. 2, p. e0211116, Feb. 2019, doi: 10.1371/journal.pone.0211116.

[11]G. Bonaccorso, *Machine Learning Algorithms*. Packt Publishing Ltd, 2017. Accessed: Jan. 12, 2023. [Online]. Available: https://books.google.com.bd/books?hl=en&lr=&id=_-ZDDwAAQBAJ&oi=fnd&pg=PP1&dq=machine+learning&ots=epkBC-Fy2C&sig=AsWG51gC9_X28Nuc_C1akXhaDPU&redir_esc=y#v=onepage&q=machine%20learning&f=false

[12]P. Moreno-Muñoz, L. Romero-Medrano, Á. Moreno, J. Herrera-López, E. Baca-García, and A. Artés-Rodríguez, “Passive detection of behavioral shifts for suicide attempt prevention,” *arXiv:2011.09848 [cs]*, Nov. 2020, doi: 10.48550/arXiv.2011.09848 Focus to learn more.

[13]E. D. Klonsky, A. M. May, and B. Y. Saffer, “Suicide, Suicide Attempts, and Suicidal Ideation,” *Annual Review of Clinical Psychology*, vol. 12, no. 1, pp. 307–330, Mar. 2016, doi: 10.1146/annurev-clinpsy-021815-093204.

[14]A. T. Beck, M. Kovacs, and A. Weissman, “Assessment of suicidal intention: The Scale of Suicide Ideation,” *Journal of Consulting and Clinical Psychology*, May 1979, doi: 10.1037/0022-006X.47.2.343.

[15]O. Oseguera, A. Rinaldi, J. Tuazon, and A. C. Cruz, “Automatic Quantification of the Veracity of Suicidal Ideation in Counseling Transcripts,” *Communications in Computer and Information Science*, pp. 473–479, 2017, doi: 10.1007/978-3-319-58750-9_66.

[16]A. E. Aladağ, S. Muderrisoglu, N. B. Akbas, O. Zahmacioglu, and H. O. Bingol, “Detecting Suicidal Ideation on Forums: Proof-of-Concept Study,” *Journal of Medical Internet Research*, vol. 20, no. 6, p. e215, Jun. 2018, doi: 10.2196/jmir.9840

[17]“Data Mining vs Machine Learning - Javatpoint,” *www.javatpoint.com*. https://www.javatpoint.com/data-mining-vs-machine-learning (accessed Jan. 12, 2023).

[18]“Classification Algorithm in Machine Learning - Javatpoint,” www.javatpoint.com. https://www.javatpoint.com/classification-algorithm-in-machine-learning (accessed Jan. 12, 2023).

[19]“Logistic Regression in Machine Learning - Javatpoint,” *www.javatpoint.com*. https://www.javatpoint.com/logistic-regression-in-machine-learning

[20]javatpoint, “Support Vector Machine (SVM) Algorithm - Javatpoint,” www.javatpoint.com. https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm

[21]javaTpoint, “Machine Learning Decision Tree Classification Algorithm - Javatpoint,” www.javatpoint.com. https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm

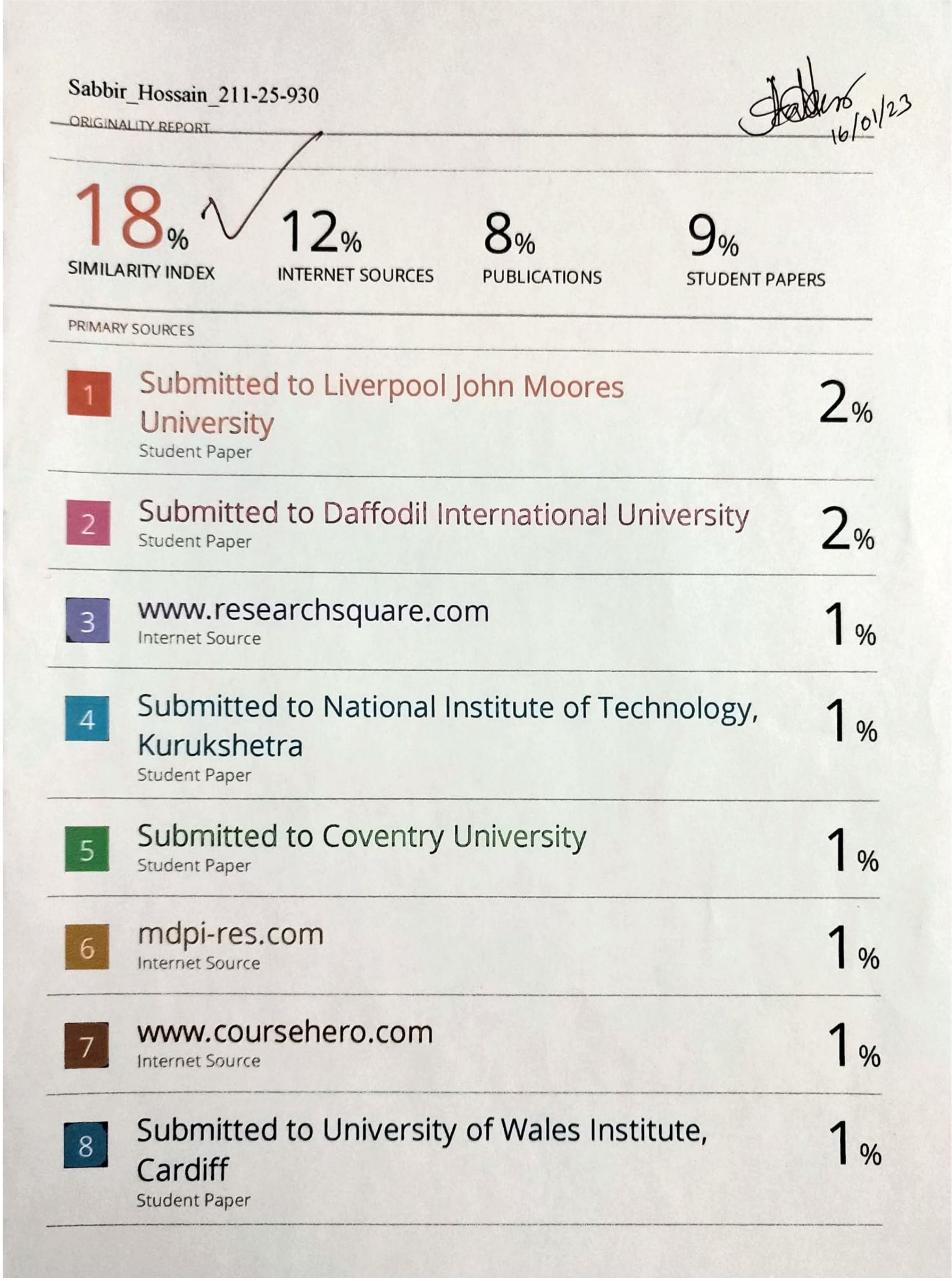
[22]“Machine Learning Random Forest Algorithm - Javatpoint,” www.javatpoint.com. https://www.javatpoint.com/machine-learning-random-forest-algorithm

[23]B. Butterfly, “Optical Neural Networks: The Future of Deep Learning?,” FindLight Blog, Oct. 26, 2022. https://www.findlight.net/blog/optical-neural-networks-the-future-of-deep-learning/ (accessed Jan. 16, 2023).

[24]JavaTpoint, “K-Nearest Neighbor(KNN) Algorithm for Machine Learning - Javatpoint,” www.javatpoint.com. https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning

[25]“Machine Learning Techniques - Javatpoint,” www.javatpoint.com. https://www.javatpoint.com/machine-learning-techniques

[26]admin, “Data Mining Techniques,” Sep. 26, 2021. https://dishcoachingcentre.com/data-mining-techniques/ (accessed Jan. 16, 2023).



Sabbir_Hossain_211-25-930

16/01/23

ORIGINALITY REPORT

18%	12%	8%	9%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Liverpool John Moores University Student Paper	2%
2	Submitted to Daffodil International University Student Paper	2%
3	www.researchsquare.com Internet Source	1%
4	Submitted to National Institute of Technology, Kurukshetra Student Paper	1%
5	Submitted to Coventry University Student Paper	1%
6	mdpi-res.com Internet Source	1%
7	www.coursehero.com Internet Source	1%
8	Submitted to University of Wales Institute, Cardiff Student Paper	1%