

TeenDepresso: An Essemble Model To Detect Depression

INTERDISCIPLINARY PROJECT

Submitted in partial fulfillment of the requirements for the award of
Bachelor of Technology in Information Technology

by

NAMAN RAJ (Reg.No-41120118)



**DEPARTMENT OF INFORMATION TECHNOLOGY
SCHOOL OF COMPUTING**

SATHYABAMA

**INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
CATEGORY-1 UNIVERSITY BY UGC
Accredited "A++" by NAAC**

12B Status by UGC |Approved by AICTE

JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI-600119

MAY 2024



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

[DEEMED TO BE UNIVERSITY]

Accredited "A++" Grade by NAAC | 12B Status by UGC | Approved by AICTE

www.sathyabama.ac.in

DEPARTMENT OF INFORMATION TECHNOLOGY

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **NAMAN RAJ (41120118)** who carried out the project entitled "**TeenDepresso: An Essemble Model To Detect Depression**" under my supervision from January 2024 to May 2024.

Internal Guide

Ms.P.Shamili M.Tech.,

Head of the Department

Dr. R. Subhashini, M.E., Ph.D.,

Submitted for Viva-voce Examination held on 08/05/2024

Internal Examiner

8/5/24
External Examiner

DECLARATION

I, **NAMAN RAJ (41120118)** hereby declare that the Project Report entitled "**TeenDepresso: An Ensemble Model To Detect Depression**" done by me under the guidance of **Ms.P.Shamili M.Tech.**, is submitted in partial fulfillment of the requirements for the award of Bachelor of Technology degree in **Information Technology**.

DATE: 6/5/2024

PLACE: CHENNAI



SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to the **Board of Management of SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully.

I am grateful to them. I convey my thanks to **Dr T.Sasikala M.E., PhD, Dean**, School of Computing, and **Dr R.Subhashini M.E., Ph.D.**, Head of the Department of Information Technology for providing me with necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Ms.P.Shamili M.Tech.**, for her valuable guidance, suggestions and constant encouragement paved the way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Information Technology** who were helpful in many ways for the completion of the project.

CERTIFICATION



Elite NPTEL Online Certification

(Funded by the MoE, Govt. of India)



This certificate is awarded to

NAMAN RAJ

for successfully completing the course

Data Science for Engineers

with a consolidated score of **69** %

Online Assignments	23.96/25	Proctored Exam	44.84/75
--------------------	----------	----------------	----------

Total number of candidates certified in this course: **2868**

Devendra Jalihal

Prof. Devendra Jalihal
Chairperson,
Centre for Outreach and Digital Education, IITM

Jan-Mar 2024

(8 week course)

Prof. Andrew Thangaraj

Prof. Andrew Thangaraj
NPTEL, Coordinator
IIT Madras



Indian Institute of Technology Madras



Roll No: NPTEL24CS53S553400636

To verify the certificate



No. of credits recommended: 2 or 3

ABSTRACT

In recent times, intensifying struggles for sustenance and education, have sharply escalated competitive tensions among individuals. This heightened competition has left many students, especially adolescents in prolonged states of stress and apprehension, due to the lack of proper guidance, trauma or other aspects leading to a significant surge in mental health issues. The recent advancement in social media like Instagram, Facebook, Twitter, etc. has made communication easy but also has become something that is offering these teenagers a fresh opportunity or method to release or express their emotions which many of their age group tend to keep to themselves. This could be a platform, activity, or resource that allows them to openly share and manage their feelings, thus serving as a beneficial outlet for their emotional well-being. They tend to share their lives through the social network and interact with their friends online with whom they feel at ease to share their thoughts. All this has greatly helped in detecting depressed individuals by analyzing their social network data. Here this study involves text-based analysis of the Twitter posts by the adolescent users to detect depression in them. This starts with the collection of textual data from Twitter posts from teenagers and then converting it into input data that can be used for Machine Learning. Several kinds of machine learning algorithms will be experimented with and finally, the best one can be declared. This depression recognition scheme refers to a method or system designed to identify signs or patterns associated with depression in individuals by utilizing data from Twitter.

TABLE OF CONTENT

Chapter No.	Title	Page No.
	ABSTRACT	v
	LIST OF ABBREVIATIONS	viii
	LIST OF TABLES	ix
	LIST OF FIGURES	x
1	INTRODUCTION	1
2	METHODOLOGY	4
2.1	DEPRESSION CLASSIFICATION	4
2.2	DEPRESSION RECOGNITION	7
2.3	RECOGNITION METHODS	9
2.4	DEPRESSION DATASET	12
3	DEPRESSION RECOGNITION OF TEENAGERS	16
3.1	RECOGNITION MODEL	16
3.2	COLLECTING DATASET & PREPROCESSING	17
3.3	FEATURE EXTRACTION	17
3.4	BUILDING THE MODEL	19
4	RESULT & DISCUSSION	25

4.1	RESULT	25
4.2	DISCUSSION	26
5	CONCLUSION	27
	REFERENCES	28
	A. SOURCE CODE	
	B. SCREENSHOTS	

LIST OF ABBREVIATIONS

X	:	Twitter
EEG	:	Electroencephalography
HRV	:	Heart Rate Variability
PET	:	Positron Emission Tomography
MRS	:	Magnetic Resonance Spectroscopy
KNN	:	K Nearest Neighbor
SVM	:	Support Vector Machine

LIST OF TABLES

TABLE No:	TABLE NAME	PAGE No:
2.1	Hierarchical diagnosis results of Baker scale	9
2.2	Hamilton scale	10
3.1	Sample size details	17
3.2	Comparison of high-frequency words	18
3.3	Example of expression classification	18
4.1	Comparison Table	25

LIST OF FIGURES

FIGURE NO	FIGURE NAME	PAGE NO
2.1	ML Model	11
2.2	Deep Learning Model	12
2.3	Sample distribution of Avid-Corpus dataset	12
2.4	Sample distribution of the DAIC-WOZ	13
2.5	CBow model	14
2.6	Deep neural network structure diagram	15
3.1	Design model design flow chart	16
3.2	Accuracy	20
3.3	Precision	20
3.4	F1 Score	20
3.5	Confusion Matrix	21
3.6	Accuracy	21
3.7	Precision	21
3.8	F1 Score	22
3.9	Confusion Matrix	22
3.10	Accuracy	23
3.11	Precision	23
3.12	Confusion Matrix	24

CHAPTER 1

INTRODUCTION

Depression is a common but serious mood disorder that affects how a person feels, thinks, and handles daily activities. Depression is something that can affect an individual's life immensely and can have a profound impact on teenagers, affecting their academic performance, social relationships, and overall quality of life. It can also increase the risk of substance abuse, self-harm, and suicidal behaviors. According to the World Health Organization (WHO), over 264 million people of all ages suffer from depression worldwide. It's increasingly recognized as a common mental health concern among young people, particularly teenagers and young adults. Depression doesn't have a single cause; it can result from a combination of genetic, biochemical, environmental, and psychological factors. Basically, in current times, medical and psychological professionals identify potential psychological disorders in individuals through the use of targeted survey questionnaires.

This project is based on participant's feelings, behaviors, and thoughts to gauge their mental health status. The scores from these surveys are used to decide if the participant is suffering from depression or not. Nonetheless, this strategy is suited primarily for individualized surveys as while this method works well on an individual level, attempting to apply it to a large population would be impractical. It would require an enormous amount of time, effort, and resources to distribute the surveys, collect the responses, and analyze the data for a large group of people.

Twitter, now X is one of the most popular social media platforms in India. Twitter is known for its real-time updates, making it an excellent source for capturing the immediate sentiments and reactions of the public to events as they happen. Twitter has millions of active users who generate a vast amount of data daily through tweets. This provides a rich dataset that reflects a wide range of opinions, emotions, and topics from diverse demographics and geographic locations Unlike some social media platforms where much of the content is shared privately, a large portion of Twitter content is public.

This accessibility makes it easier for researchers to collect and analyze data without the same privacy and access constraints present on other platforms. Tweets are limited in length, encouraging users to express their thoughts concisely. This brevity can simplify the process of sentiment analysis by reducing the complexity of the text to be analyzed, compared to longer content formats. Twitter's use of hashtags and its ability to highlight trending topics help researchers quickly identify and analyze conversations around specific themes or events. This feature simplifies the aggregation of related tweets for sentiment analysis. Twitter's 280-character limit (formerly 140 characters) encourages users to distil their thoughts into short, direct messages. This constraint fosters brevity and encourages users to focus on expressing key emotions or opinions succinctly. The multimedia support enables users to convey emotions more vividly and creatively, adding depth to their expressions. While Twitter is a public platform by default, users have control over their privacy settings. They can choose to protect their tweets or maintain anonymity by using pseudonyms. This flexibility can encourage more authentic and open emotional expression. analyze conversations around specific themes or events. This feature simplifies the aggregation of related tweets for sentiment analysis. Twitter's 280-character limit (formerly 140 characters) encourages users to distil their thoughts into short, direct messages. This constraint fosters brevity and encourages users to focus on expressing key emotions or opinions succinctly. The multimedia support enables users to convey emotions more vividly and creatively, adding depth to their expressions. While Twitter is a public platform by default, users have control over their privacy settings. They can choose to protect their tweets or maintain anonymity by using pseudonyms. This flexibility can encourage more authentic and open emotional expression.

For this project, I gathered data from adolescents who are active users of Twitter and are in their first year of undergraduate studies. Our aim was to gain insights into their experiences with depression during this transitional phase in their lives. The focus of our study was on teenage Twitter users who had just started their first year of college. We wanted to explore how they were coping with and experiencing depression

during this critical period of transition. The data utilized in this study was derived from a survey that was administered to these undergraduate students. This survey was designed specifically to gather information about their experiences and perspectives related to depression. Thear then collected data was pre-processed and used with the " SVM" algorithm for training the model. Here the potential for identifying individuals with depression through their activity on the Twitter social network is explored from a theoretical standpoint. The feasibility of detecting depression among users based on data from the Twitter social network is explored in theory. The current research focused on analyzing users' mental health using social network behaviors and published texts tend to prioritize platforms like Instagram and Facebook. It's observed that college students who may be experiencing depression often exhibit low interest in social activities. As a result, they typically have fewer connections and engage less frequently with friends on Instagram and Facebook, resulting in relatively inactive profiles.

In contrast, Twitter boasts a large user base and offers anonymity. This anonymity may encourage individuals, including those suffering from depression, to express their genuine thoughts and emotions more openly. Consequently, Twitter provides a more effective reflection of a user's daily life and mental state, offering clearer insights into depression and other mental health issues. Here, for this analysis, I am going to take only the posts in the English language into consideration

CHAPTER 2

METHODOLOGY

2.1 Depression Classification:

Depression can be classified into the following types.

2.1.1 Types of Depression:

In contrast, Twitter boasts a large user base and offers anonymity. This anonymity may encourage individuals, including those suffering from depression, to express their genuine thoughts and emotions more openly. Consequently, Twitter provides a more effective reflection of a user's daily life and mental state, offering clearer insights into depression and other mental health issues. Here, for this analysis, we are going to consider only the posts in the English language.

Mild Depression:

Mild depression, also known as dysthymia or persistent depressive disorder, is characterized by the presence of depressive symptoms that are generally milder in intensity compared to moderate or severe depression. Individuals with mild depression may experience symptoms such as low mood, loss of interest or pleasure in activities, fatigue Depression is a common but serious mood disorder that affects how a person feels, thinks, and handles daily activities. Depression is something that can affect an individual's life immensely and can have a profound impact on teenagers, affecting their academic performance, social relationships, and overall quality of life.

Moderate depression involves more pronounced symptoms that typically interfere with daily functioning to a greater extent than mild depression. In addition to experiencing the symptoms mentioned above, individuals with moderate depression may have difficulty carrying out routine tasks, such as work or school responsibilities, household chores, or social activities.

Severe Depression:

Severe depression is the most debilitating form of the condition, characterized by intense and pervasive symptoms that significantly impair all aspects of an individual's life. In addition to the symptoms present in mild and moderate depression, severe depression may involve suicidal thoughts or behaviors, psychomotor agitation or retardation, persistent feelings of hopelessness or despair, and marked impairment in cognitive functioning, such as difficulty making decisions or concentrating. Severe depression poses a serious risk to the individual's safety and requires urgent intervention, including intensive therapy, medication, hospitalization, and/or other forms of support.

2.1.2 Divided By Number of Attacks:

Depression can also be classified into different categories based on the recurrence or frequency of depressive episodes.

Single Episode Depression:

This category includes individuals who have experienced only one episode of major depression in their lifetime. The episode may be mild, moderate, or severe, but it is characterized by a distinct period of depressive symptoms that lasts for at least two weeks and significantly impairs functioning. After the episode resolves, the individual may not experience another depressive episode.

Recurrent Depression:

Recurrent depression involves multiple episodes of major depression separated by periods of partial or full remission. These episodes may vary in severity and duration. Recurrent depression is further classified into:

- ***Two Episodes:*** Individuals who have experienced two or more distinct episodes of major depression
- ***Multiple Episodes (more than two):*** Individuals who have a history of three or more distinct episodes of major depression. The frequency and pattern of recurrence can vary widely among individuals.

Chronic Depression:

Chronic depression, also known as persistent depressive disorder (PDD) or dysthymia, involves a long-term, low-grade depression that persists for at least two years in adults (one year in children or adolescents). While the symptoms of chronic depression may be less severe than those of major depression.

Cyclothymic Disorder:

Cyclothymic disorder is a milder form of bipolar disorder characterized by recurrent periods of hypomanic symptoms alternating with periods of depressive symptoms. These mood fluctuations are less severe than those seen in bipolar disorder but still have a significant impact on functioning.

2.1.3 Divided By Individual Age Level:***Childhood Depression:***

Childhood depression can often be challenging to diagnose due to differences in symptom presentation compared to adults. Symptoms may include irritability, clinginess, somatic complaints (such as headaches or stomachaches), social withdrawal, decreased interest in play activities, changes in appetite or sleep patterns, and academic or behavioral problems at school. Childhood depression may also manifest as separation anxiety or oppositional behavior.

Adolescent Depression:

Adolescence is a period of significant physical, emotional, and social development, making it a vulnerable time for the onset of depression. Symptoms of adolescent depression may include persistent sadness, irritability, anger, loss of interest in activities, social withdrawal, changes in sleep and appetite, academic decline, self-harm behaviors, substance abuse, and thoughts of death or suicide. Adolescents may also experience increased conflict with parents or authority figures and have difficulty expressing their emotions or seeking help.

Adult Depression:

Depression in adults often presents with classic symptoms such as persistent sadness, loss of interest or pleasure in activities, fatigue, changes in appetite or

weight, sleep disturbances, feelings of worthlessness or guilt, difficulty concentrating, and thoughts of death or suicide. Adults may also experience psychosomatic symptoms, such as headaches or digestive issues, and may engage in behaviors such as excessive working or substance abuse as coping mechanism.

Elderly Depression:

Depression in older adults can be underdiagnosed and undertreated, often mistaken for normal ageing or other medical conditions. Symptoms may include persistent sadness, fatigue, loss of interest in previously enjoyed activities, social withdrawal, cognitive impairment, sleep disturbances, physical complaints, and thoughts of death or suicide. Elderly individuals may also have coexisting medical conditions, sensory impairments, mobility issues, or bereavement that contribute to depression. Treatment considerations for elderly depression may include managing comorbid medical conditions, addressing social isolation, and ensuring accessibility to mental health services.

2.2 Depression Recognition:

2.1.4 Physiological Signals:

Electroencephalography (EEG): EEG is a valuable tool for studying brain activity non-invasively. Research has identified specific EEG patterns associated with depression, such as increased frontal alpha asymmetry and reduced connectivity in brain networks involved in emotion regulation.

Heart Rate Variability (HRV): HRV reflects the autonomic nervous system's regulation of heart rate and has been linked to emotional regulation and stress response. Individuals with depression may exhibit reduced HRV, indicating dysregulation of the autonomic nervous system.

Skin Conductance: Skin conductance, also known as electrodermal activity, measures changes in sweat gland activity in response to emotional stimuli. Studies have found altered skin conductance responses in individuals with depression, suggesting aberrant emotional processing.

Eye Movement Patterns: Eye tracking technology can detect subtle changes in gaze behavior, such as increased fixation duration on negative stimuli or avoidance of eye contact, which may serve as markers of depression.

Changes in Speech Patterns: Natural language processing techniques can analyze speech patterns for features such as speech rate, intonation, and word choice indicative of depression.

Social Withdrawal: Social media data and smartphone usage patterns can provide insights into social withdrawal and isolation, which are common features of depression.

Changes in Activity Levels: Wearable activity trackers and smartphone sensors can monitor physical activity levels and detect changes associated with depressive symptoms, such as decreased movement or sedentary behavior.

Sleep Disturbances: Actigraphy and sleep tracking devices can assess sleep duration, quality, and disruptions characteristic of depression, such as insomnia or fragmented sleep.

2.1.5 Cognitive Markers:

Impaired Concentration: Cognitive tests and computerized tasks can assess attention and concentration abilities, identifying impairments commonly observed in depression.

Negative Thought Patterns: Natural language processing algorithms can analyze text data, such as journal entries or social media posts, to identify negative thought content and rumination associated with depression.

Impaired Executive Functioning: Neuropsychological assessments and cognitive tests can evaluate executive functions such as planning, problem-solving, and decision-making, which may be impaired in depression.

2.1.6 Biological Markers:

Neurotransmitter Levels: Neuroimaging techniques such as positron emission tomography (PET) and magnetic resonance spectroscopy (MRS) can measure neurotransmitter levels in the brain, providing insights into the neurochemical basis of depression.

Neuroendocrine Functioning: Salivary cortisol assays and dexamethasone suppression tests can assess HPA axis function and cortisol dysregulation, which are associated with depression.

Inflammatory Markers: Blood tests can measure levels of inflammatory markers such as interleukin-6 (IL-6) and tumor necrosis factor-alpha (TNF-alpha), which have been implicated in the pathophysiology of depression.

2.2 Depression Recognition Methods:

2.2.1 Traditional depression recognition

Traditional diagnostic tools include interview assessments, the Hamilton Depression Rating Scale, and the Baker Depression Scale. These scales and the severity of behavioral symptoms can give a certain grade of scores to patients with depression. The diagnosis using this method is extremely complicated. It depends heavily on the patient's ability, patience, willingness, and whether the patient is honest and serious when communicating with them. However, according to the definition of depression, their thinking and motivation are impaired, so diagnostic information takes a long time to collect. Patient training, practice and validation are needed to produce acceptable results.

TABLE 2.1 : Hierarchical diagnosis results of Baker scale

Measurement score	Analysis of measurement results	Depression treatment recommendations
0-4	No Depression	None
5-9	Mild depression	Wait for observation
10-14	Moderate depression	Counselling or mild

table 2.1 shows the classification of the most commonly used Baker scale results, showing the severity of depression in five levels. table 2.2 shows the Hamilton depression scale commonly used in Europe.

TABLE 2.2 Hamilton scale

	None	Mild	Moderate	Severe	Very severe
1. Anxiety	0	1	2	3	4
2. tensivity	0	1	2	3	4
3. funk	0	1	2	3	4
4. insomnia	0	1	2	3	4
5. Memory or attention disorder	0	1	2	3	4
6. Depression	0	1	2	3	4
7. Muscle system symptoms	0	1	2	3	4
8. Sensory system symptoms	0	1	2	3	4
9. Cardiovascular system symptoms	0	1	2	3	4
10. Respiratory symptoms	0	1	2	3	4
11. Gastrointestinal symptoms	0	1	2	3	4
12. Genitourinary symptoms	0	1	2	3	4
13. Autonomic symptoms	0	1	2	3	4
14. Behavior during the talks	0	1	2	3	4

Diagnosing depression in this way is very difficult at the primary treatment stage. The diagnosis of many depressions can also be very wide. Moreover, the results obtained by such classification and classification are very general, can only the people that they are suffering from high or low level of depression.

2.2.2 Recognition model based on machine learning

At present, the commonly used depression recognition models based on machine learning are Decision Trees, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Naive Bayesian Model (NBM), Random Forest) and other machine-learning models. The flowchart of depression recognition based on machine learning is shown in figure 2.1.

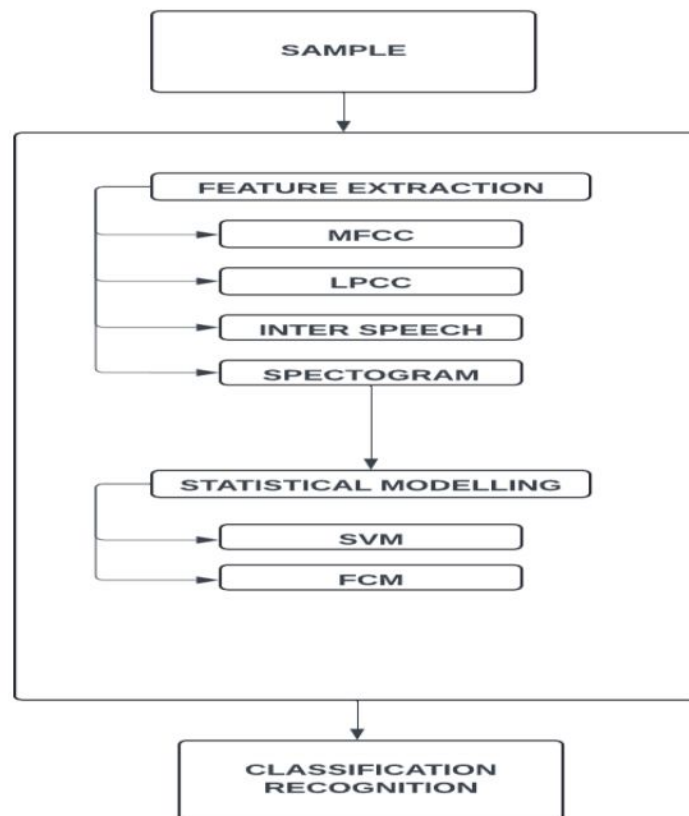


FIGURE 2.1 ML Model

2.3.2 Recognition Model Based on Deep Learning:

As data gradually becomes more dimensional and diverse, data processing becomes more and more important. Deep neural networks are powerful processing data and are powerful tools for processing complex data. Pitts simulates the operation of the human brain to store and process complex information. It can process data in parallel and dig more useful information from the data. Convolutional neural network is one of the commonly used depression recognition models. figure 2.2 shows a flowchart of depression recognition based on a deep learning.

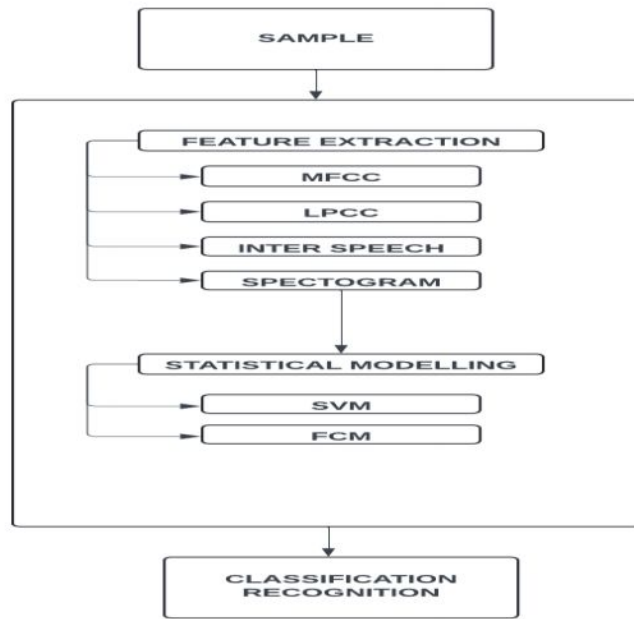


FIGURE 2.2 DL Model

2.3 Depression Dataset:

1. Avid-Corpus Depression Speech Dataset

As data gradually becomes more dimensional and diverse, data processing becomes more and more important. Deep neural networks are powerful in processing data and are powerful tools for processing complex data. Pitts simulates the operation of the human brain to store and process complex information. It can process data in parallel and dig more useful information from the data. Convolutional neural network is one of the commonly used depression recognition models. figure 2.3 shows a flowchart of depression recognition based on a deep convolutional neural network.

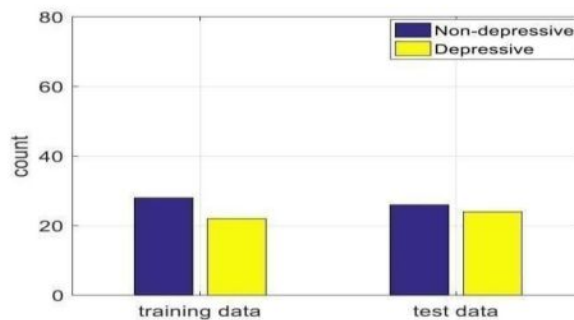


FIGURE 2.3 Sample distribution of Avid-Corpus dataset

2. DAIC-WOZ Depression Speech Data Set

The DAIC-WOZ data set is mainly for audio, video, and questionnaire feedback data collected for the diagnosis of anxiety, depression, and other psychological diseases. The distribution of depression and non-depression individuals in this data set is shown in figure 2.4.

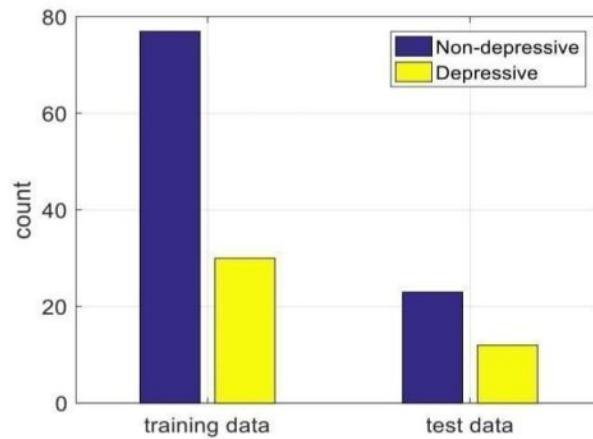


FIGURE 2.4 Sample distribution of the DAIC-WOZ dataset

3. Text data in social networks

Contemporary college students like social networks such as Twitter, and post, repost, and comment on various information on Weibo. It is possible to analyze and identify whether the user has a depression tendency from the text information. How to translate the information in this article into input vectors available to the classifier. There are generally representation models in this paper, which are vector space model, theme model and Word Embedding. In Word Embedding, the proposal of word2vec enables text to be converted into vectors like pictures and also promotes the development of deep learning in the field of natural language processing. word2vec has two models-skip-gram model and the Cow model. The structure of the two models is shown in figures 2.5.

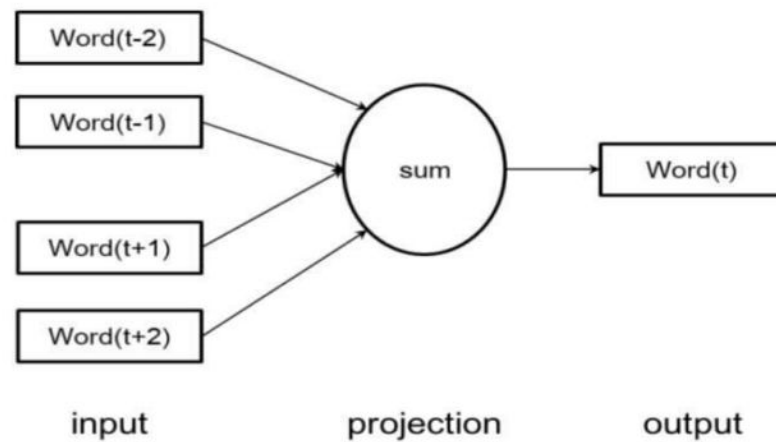


FIGURE 2.5 CBow model

CBow is a model that predicts the current word based on the context. It has the same structure as the ordinary feed-forward neural network and uses a local sliding window mechanism. The selected window size in figure 2.5 is 4, where $w(t)$ is the word vector at position t , and it is also the central word to be predicted. The model has three layers: one is the input layer. The context word vector $w(t-2)$, $w(t-1)$, $w(t+1)$, and $w(t+2)$ of the word at the input position is represented by a distributed representation method. In the middle is the projection layer, which can superimpose the input word vector. The main idea of the projection layer is the Huffman tree. The third time is the output layer to get a vector of words at position.

4. Deep neural network

In this study, deep neural networks are used for feature extraction. In pattern recognition, features are mainly used to describe and represent the things studied. The features extracted by the algorithm from the data directly affect the performance of the model. How to effectively extract features that conform to the data distribution structure has always been a hot issue in this field. Traditional feature extraction methods mostly select features based on different tasks and data designs, such as Gabor features, SIFT, local binary patterns, etc.

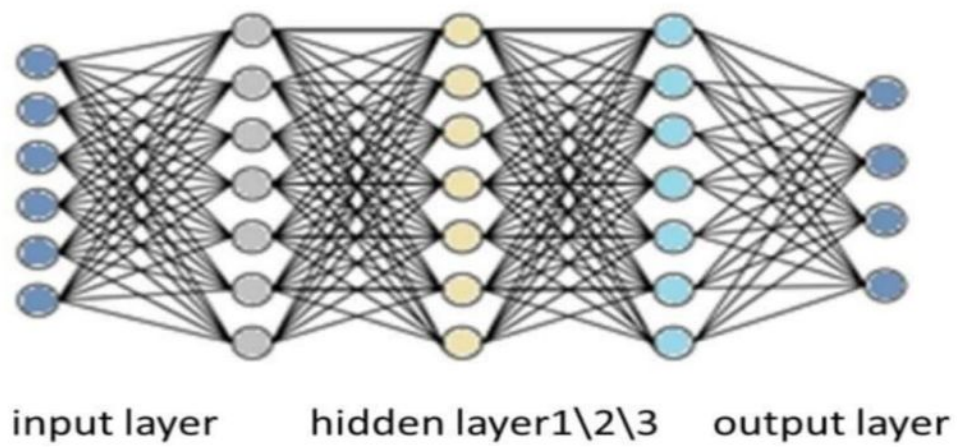


Figure 2.6 Deep neural network structure diagram

However, designing a good feature is not simple, especially in the case of a large amount of data, it takes time and effort. How to find a method that can automatically learn feature description is very critical for this research as shown in figures 2.6.

CHAPTER 3

DEPRESSION RECOGNITION

3. DEPRESSION RECOGNITION OF TEENAGERS

3.1 Recognition Model

The functional modules involved in the depression prediction model are shown in figure 3.1, which can be divided into three layers as a whole: the acquisition of raw data, data preprocessing, and training recognition model. Acquisition and preprocessing of raw data: This part is the basis of the entire technical framework. The user's raw data on social media is mainly obtained through the Internet. Then through step-by-step processing, the user network behavior characteristics are finally stored in the database.

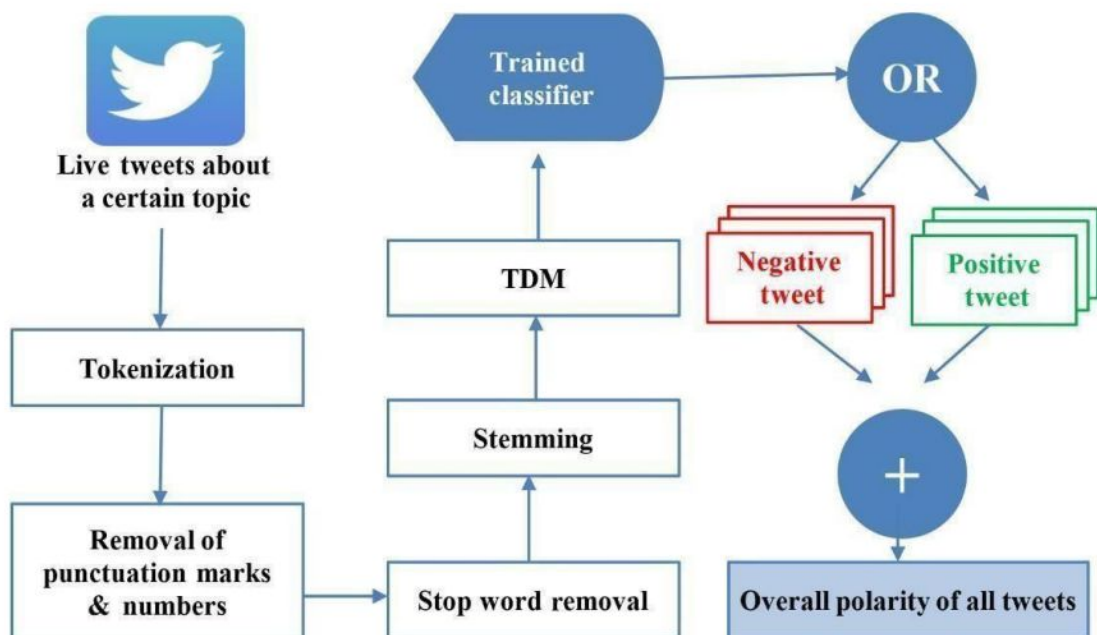


FIGURE 3.1 Design model design flow chart

3.2 Obtaining Data Sets and Preprocessing

Obtain the user data required to build a depression recognition model and perform preprocessing such as de-redundancy, cleaning, and sorting of these data. The data source needed for the depression recognition model constructed in this paper is the Internet behavior characteristics of Internet users on social media.

Specifically, we use the third-party interface provided by Sina Weibo and our own crawler software to obtain college student Weibo users' online behavior data on Sina. Then combined with the network behavior system.

First, we took a dataset of 446 samples from the social media platform "Twitter" now "X" related to the sentiments of the people. After observing the dataset, I labelled the data set as a depressed-prone sample and a normal-prone sample. After applying the pre-processing, we get to number the number depressed and normal sentiments. In table 3.1 it can be easily seen.

TABLE 3.1 Sample size details:

Total sample	Depression-prone samples	Number of normal users
446	209	237

3.3 Feature analysis and extraction

Analysis of high-frequency words on Twitter According to the research results in the literature, it is known that depression patients mostly use the first person when speaking. In addition, depression patients like to post and repost negative and negative messages. Therefore, this paper constructs input data by counting the frequency of specific words on Twitter. Specific words generally include 6 levels: degree words, negative evaluation words, negative emotional words, positive evaluation words, positive emotional words, and propositional words. Some words belong to multiple general categories at the same time, for example, the word "very" appears in the degree level, positive evaluation words, and other categories. To simplify the problem, this article ignores the part of speech. By removing duplication, each word appears only once in the vocabulary. Table 3.2 shows the comparison of high-frequency words sent by depression patients and normal people.

TABLE 3.2 Comparison of high-frequency words

Number	Normal user	Depression users
1	Excited	unable
2	good	weight
3	party	darkness
4	great	sadness
5	family	emptiness
6	love	never
7	team	escape
8	progress	pain
9	thankful	trapped
10	enjoyed	sleep

In addition to text content, Sina Weibo also supports pictures, videos, music, emojis, and more. Contents such as pictures and videos are too large and difficult to obtain. This study does not involve this part of the data. The Weibo content data obtained in this article includes emojis. For example, the symbol "candle" is represented as "[candle]" in the text. Emoji is a very popular form of expression on the Internet. It refers to static or dynamic pictures of people's expressions contained in text. For example, the expression "dagger" on Sina Weibo is usually used to express anger. Generally, expressions can be divided into three categories: positive, negative, and neutral. Examples of expression classification are shown in table 3.4

TABLE 3.3 Example of expression classification:

Number	Expression	Character
1	[Haha]	Positive
2	[tears]	Neutral
3	[love you]	Positive
4	[anger]	Negative

Through the analysis of the Twitter data of the experimental sample throughout 2024, we found that college students with depression use negative and neutral expressions more frequently, and normal college students use positive and neutral expressions more often. The comparison of the two types of college student users using positive, negative and neutral expressions.

3.1 Building the Model

3.1.1 Naive Bayes Model

The Naive Bayes algorithm is a simple yet powerful probabilistic classification technique widely used in machine learning. It is based on the Bayes theorem with the assumption of independence between features. Bayes' Theorem provides away to calculate the probability of a hypothesis given some evidence. It is expressed as:

$$P(A/B)= \frac{P(B/A) \times P(A)}{P(B)}$$

Where:

- $P(A/B)$ is the posterior probability of hypothesis A given the evidence B.
- $P(B/A)$ is the likelihood of evidence B given the hypothesis A.
- $P(A)$ is the prior probability of hypothesis A.
- $P(B)$ is the total probability of the evidence.

In the context of classification, Bayes' theorem is used to predict the probability of each class given the input features. The Naive Bayes classifier makes the assumption that the features are conditionally independent given the class label. The class with the highest posterior probability is predicted. The parameters of the Naive Bayes model include prior probabilities and conditional probabilities. These parameters are estimated from the training data using maximum likelihood estimation (MLE).

Strengths: Naive Bayes models are simple, efficient, and scalable to large datasets. They perform well in text classification and other tasks with high- dimensional data.

Weaknesses: The strong assumption of feature independence may not hold in some real-world scenarios. Naive Bayes may perform poorly if this assumption is The Naive Bayes algorithm offers a straightforward yet effective approach to classification tasks, particularly in text and document analysis. Its simplicity, efficiency, and scalability make it a valuable tool in various machine-learning applications.

ACCURACY OBTAINED

The ACCURACY obtained from the Naive Bayes Algorithm is **88.06%**

```
from sklearn.naive_bayes import ComplementNB
from sklearn.metrics import classification_report, confusion_matrix
CNB = ComplementNB()
CNB.fit(X_train, y_train)

from sklearn import metrics
predicted = CNB.predict(X_test)
accuracy_score = metrics.accuracy_score(predicted, y_test)
print('Naive Bayes model accuracy is', str('{:04.2f}'.format(accuracy_score*100)) + '%')
```

FIGURE 3.2 Accuracy

PRECISION OBTAINED

The PRECISION obtained from the Naive Bayes Algorithm is **90.14%**

```
from sklearn.naive_bayes import ComplementNB
from sklearn.metrics import precision_score

CNB = ComplementNB()
CNB.fit(X_train, y_train)

predicted = CNB.predict(X_test)
precision = precision_score(y_test, predicted)

print('ComplementNB model precision is', str('{:04.2f}'.format(precision * 100)) + '%')
```

FIGURE 3.3 Precision

F1 SCORE OBTAINED

The F1 SCORE obtained from the Naive Bayes Algorithm **88.98%**

```
from sklearn.metrics import f1_score
f1 = f1_score(y_test, predicted)
print('F1 Score is', str('{:04.2f}'.format(f1 * 100)) + '%')
```

FIGURE 3.4 F1 Score

CONFUSION MATRIX OBTAINED

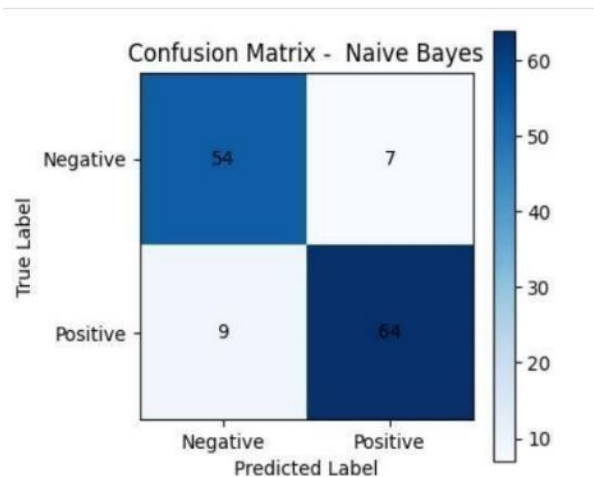


FIGURE 3.5 Confusion Matrix

3.1.2 KNN MODEL

KNN is a simple and intuitive non-parametric classification algorithm that classifies new instances based on the majority class of their K nearest neighbors in the feature space. KNN uses distance metrics (e.g., Euclidean distance, Manhattan distance) to measure the similarity or distance between data points in the feature space.

ACCURACY OBTAINED

The ACCURACY obtained from the KNN Algorithm is **67.17%**

```
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix
knn = KNeighborsClassifier(n_neighbors=5)
knn.fit(X_train, y_train)
predicted_knn = knn.predict(X_test)
accuracy_knn = accuracy_score(y_test, predicted_knn)
print('KNN model accuracy is', str('{:04.2f}'.format(accuracy_knn * 100)) + '%')
```

FIGURE 3.6 Accuracy

PRECISION OBTAINED

The PRECISION obtained from the KNN Algorithm is **80.85%**

```
from sklearn.metrics import precision_score
precision_knn = precision_score(y_test, predicted_knn)
print('KNN model precision is', str('{:04.2f}'.format(precision_knn * 100)) + '%')
```

FIGURE 3.7 Precision

F1 SCORE OBTAINED

The F1 SCORE obtained from the KNN Algorithm is **88.98%**

```
from sklearn.metrics import f1_score
f1score_knn = f1_score(y_test, predicted_knn)
print('KNN model F1score is', str('{:04.2f}'.format(f1score_knn * 100)) + '%')
```

FIGURE 3.8 F1 Score

CONFUSION MATRIX OBTAINED

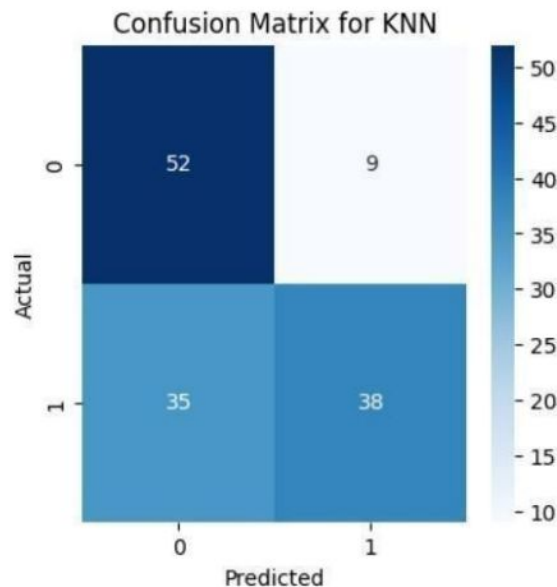


FIGURE 3.9 Confusion Matrix

3.1.3 SVM MODEL

Support Vector Machine is a powerful machine learning algorithm used for linear or nonlinear classification, regression, and even outlier detection tasks. SVMs can be used for a variety of tasks, such as text classification, image classification, spam detection, handwriting identification, gene expression analysis, face detection, and anomaly detection. SVMs are adaptable and efficient in a variety of applications because they can manage high-dimensional data and nonlinear relationships.

ACCURACY OBTAINED

The ACCURACY obtained from the **SVM** Algorithm is **90.30%**

```
from sklearn.svm import SVC
svm = SVC(kernel='linear')
svm.fit(X_train, y_train)
accuracy_score_svm = svm.score(X_test, y_test)
print('SVM model accuracy is', str('{:04.2f}'.format(accuracy_score_svm*100)) + '%')
```

FIGURE 3.10 Accuracy

PRECISION OBTAINED

The PRECISION obtained from the **SVM** Algorithm is **94.12%**

```
from sklearn.metrics import precision_score

# Predict the labels for the test set
predicted_svm = svm.predict(X_test)

# Calculate precision score for the SVM classifier
precision_svm = precision_score(y_test, predicted_svm)

print('Precision score for SVM classifier is', str('{:04.2f}'.format(precision_svm * 100)) + '%')
```

FIGURE 3.11 Precision

F1 SCORE OBTAINED

The F1 SCORE obtained from the **SVM** Algorithm is **90.78%**

```
from sklearn.svm import SVC
from sklearn.metrics import f1_score
svm_classifier = SVC(kernel='linear', C=1.0, random_state=42)
svm_classifier.fit(X_train, y_train)
y_pred = svm_classifier.predict(X_test)
f1score = f1_score(y_test, y_pred)
f1score_percentage = f1score * 100
print("F1 Score for SVM Classifier:", "{:.2f}%".format(f1score_percentage))
```

FIGURE 3.12 F1 Score

CONFUSION MATRIX OBTAINED

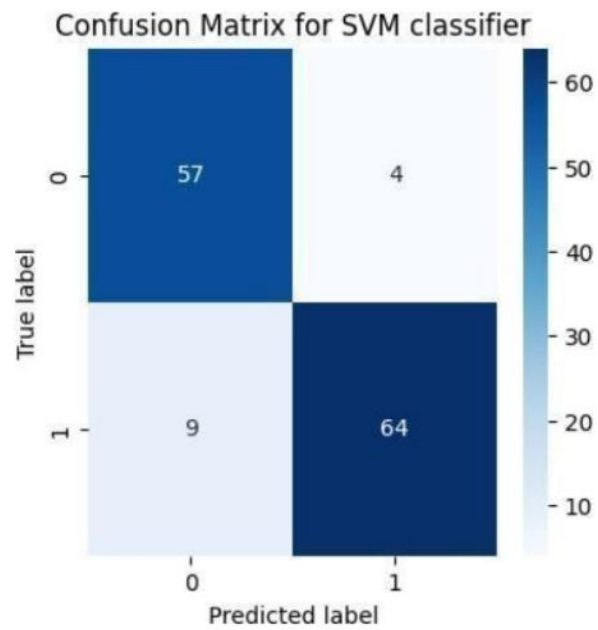


FIGURE 3.13 Confusion Matrix

CHAPTER 4

RESULT & DICSUSSION

4.1 RESULT:

For the above data set, I have used three algorithms for doing the classification. These three algorithms are Naive Bayes, KNN and SVM. I have taken a few parameters on which I am doing the model comparison. The parameters are accuracy, precision, and F1 score. The comparison is shown in the table 4.1.

TABLE 4.1 Comparison Table:

Model	ACCURACY	PRECISION	F1 SCORE
Naive Bayes	88.06%	90.14%	88.98%
KNN	67.17%	80.85%	88.98%
SVM	90.30%	94.12%	90.78%

The above table shows the percentage of different algorithms used for the model.

4.1.1 IDENTIFYING BEST ALGORITHM

From the above table, SVM is the best algorithm for the sentiment analysis of the dataset obtained front he Twitter. It has more accuracy and precision. The advantages of SVM lie in its capability to handle both linear and non-linear classification problems, capture complex decision boundaries, and generalize well to unseen data. This robust performance makes SVM a compelling choice for this specific dataset and classification problem.

4.2 DISCUSSION

The sentiment analysis algorithms—K-Nearest Neighbors, Support Vector Machines, and Naive Bayes—each offer unique strengths in detecting depression-related sentiment in text data. KNN relies on similarity metrics, SVM finds optimal hyperplanes, and Naive Bayes assumes feature independence. While KNN shows competitive performance, it may be sensitive to noise. SVM performs well with complex data but can be computationally expensive. Naive Bayes is simple and efficient but may struggle with nuanced patterns.

Performance is influenced by feature representation, dataset quality, and hyperparameter selection. Future research could explore ensemble methods, advanced feature extraction techniques, and longitudinal studies for enhanced accuracy and scalability.

Further analysis reveals the importance of considering the specific characteristics of the dataset and problem domain when selecting an appropriate algorithm. While SVM emerged as the top performer in this experiment, it is essential to explore hyperparameter tuning and optimization techniques to maximize its performance. Additionally, future research could delve into ensemble methods, neural networks, or feature engineering to further enhance classification performance. In conclusion, based on the experimental findings and performance evaluation, the Support Vector Machine stands out as the most suitable algorithm for the given classification task. Its superior accuracy and precision make it a compelling choice for practical applications, reaffirming its effectiveness in this particular context.

CHAPTER 5

CONCLUSION

After conducting thorough experimentation and evaluation of three machine learning algorithms—Naive Bayes, K-Nearest Neighbors, and Support Vector Machine (SVM)—on the dataset, it is evident that the Support Vector Machine outperforms the other algorithms in terms of accuracy and precision. Across various evaluation metrics, SVM consistently demonstrates superior performance compared to Naive Bayes and KNN. This finding underscores the effectiveness of SVM in handling the classification task at hand. The advantages of SVM lie in its capability to handle both linear and non-linear classification problems, capture complex decision boundaries, and generalize well to unseen data. This robust performance makes SVM a compelling choice for this specific dataset and classification problem.

Further analysis reveals the importance of considering the specific characteristics of the dataset and problem domain when selecting an appropriate algorithm. While SVM emerged as the top performer in this experiment, it is essential to explore hyperparameter tuning and optimization techniques to maximize its performance. Additionally, future research could delve into ensemble methods, neural networks, or feature engineering to further enhance classification performance. In conclusion, based on the experimental findings and performance evaluation, the Support Vector Machine (SVM) stands out as the most suitable algorithm for the given classification task. Its superior accuracy and precision make it a compelling choice for practical applications, reaffirming its effectiveness in this particular context.

REFERENCES

1. Ding, Xuemei Chen, Qiming Fu, Shan Zhong, "A depression recognition method for college students using deep integrated support vector algorithm" Publisher: IEEE, 2023. DOI: 10.1109/ACCESS.2020.2987523.
2. Min Kang, Seohyun Oh, KyoungSu Oh, Seokhwan Kang, Youngho Lee, "The deep learning method for predict beck's depression", Publisher: IEEE, 2022 DOI: 10.1109/ICTC52510.2021.9620922
3. Siddharth Prabhudessai, Apurva Mahaska, Manvi Parmar, Sumedha Bhagwat, "Depression Learning: Study Analysis" , Publisher: IEEE, 2023 DOI: 10.1109/CSNT51715.2021.9509707
4. Zyan Imtinan Uddin, Kaniz Fatema, Pranab Kumar Dhar, "Depression Risk Prediction among Tech Employees", Publisher: IEEE, 2024 7111. DOI: 10.1109/WIECON-ECE52138.2020.9397999
5. Zhang, J., & Zong, M. (2017), "Natural Language Processing and Chinese Computing", First CCF Conference, NLPCC 2012. Springer. DOI: 10.1109/ACCESS.2020.2987523

APPENDICES

A. SOURCE CODE:

```
import pandas as pd
import matplotlib.pyplot as plt
import plotly.express as px

from wordcloud import WordCloud
import nltk
import re
import string
from nltk.corpus import stopwords
nltk.download('punkt')
nltk.download('stopwords')
from nltk.tokenize import word_tokenize
from nltk.stem import WordNetLemmatizer

stop_words = stopwords.words()
```

```
from collections import Counter
cnt = Counter()
for text in dt["no_sw"].values:
    for word in text.split():
        cnt[word] += 1
cnt.most_common(10)
temp = pd.DataFrame(cnt.most_common(10))
temp.columns = ['word', 'count']
temp
```

```
import plotly.express as px

# Assuming 'temp' DataFrame contains columns 'word' and 'count'
fig = px.bar(temp, x="count", y="word", title='Common Words in Text', orientation='h',
             width=700, height=700, color="word")

fig.show()
```

```
FREQWORDS = set([w for (w, wc) in cnt.most_common(10)])

def remove_freqwords(text):
    """custom function to remove the frequent words"""
    return " ".join([word for word in str(text).split() if word not in FREQWORDS])

dt["wo_stopfreq"] = dt["no_sw"].apply(lambda text: remove_freqwords(text))
dt.head()
```

```
from sklearn.model_selection import train_test_split
X=text_counts
y=nb['sentiment']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.30,random_state=50)
```

B. SCREENSHOTS:

