Deciphering *Depression*

Depression Detection using Twitter Dataset Implementing Machine Learning Techniques

Manoshika Catherine S J

Introduction

In today's digital age, where social media and the internet play integral roles in the lives of countless individuals, the ability to discern emotional states holds significant importance, especially for the 300 million people affected by psychological issues. To address this pressing concern,This study focuses on analyzing textbased depression using machine learning techniques by employing a Support Vector MachineLinear SVM, Decision tree, Random Forest, Logistic Regression, and Naive Bayes algorithms on the Sentiment140 dataset with 1.6 million tweets

The Aim of this Project is to collect the twitter dataset, and to perform Feature Engineering to identify the features that are important for the classification. To identify the words that are mostly used by a depressive person and a non-depressive person, and to use various Machine learning Algorithms that are most suitable to classify the depressive person and a non-depressive person.

The overview of the Entire Project aims to forecast depression detection as an online web media post by concentrating on emotional process, linguistic foundation, and temporal features. Feature Engineering has to be performed to test and train the data. The several classifiers, including Linear Support Vector Machines, Logistic Regression, Random Forests, Decision Trees, the Naive Bayes method.

Related Works

STATISTICAL TECHNIQUES AND SENTIMENTAL ANALYSIS

Depression can be looked at by a number of statistical techniques using a machine learning algorithm to accurately identify a person's depression condition.

The analysis of the Twitter data can be done focused on identifying emotions caused by psychological disorders or mental health issues.

01

N-GRAM LANGUAGE MODELLING

Analysis combined with emotional feature analysis using N-gram language modelling is to gauge anxiety levels built a classifier to recognise clinical depression in people by studying behavioural features using a Twitter database

02

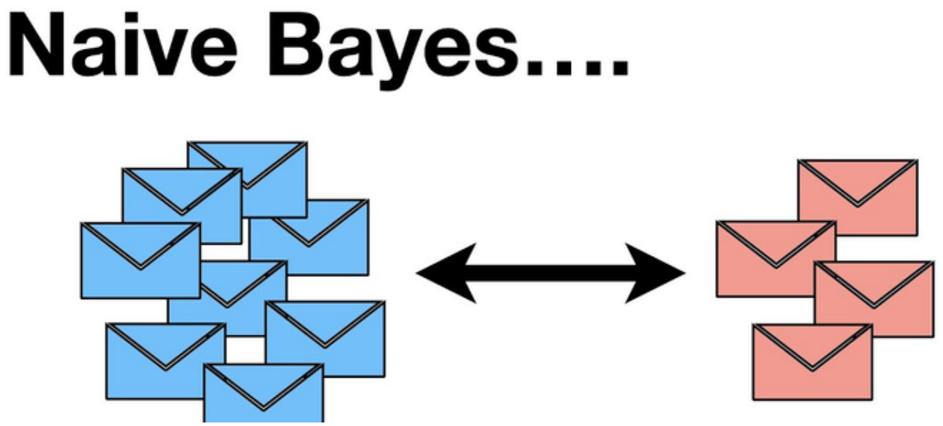
LINGUISTIC CHARACTERISTICS

Depression can be predicted using linguistic characteristics. For the training decision list, they utilized the Twitter dataset to determine the depression.

03

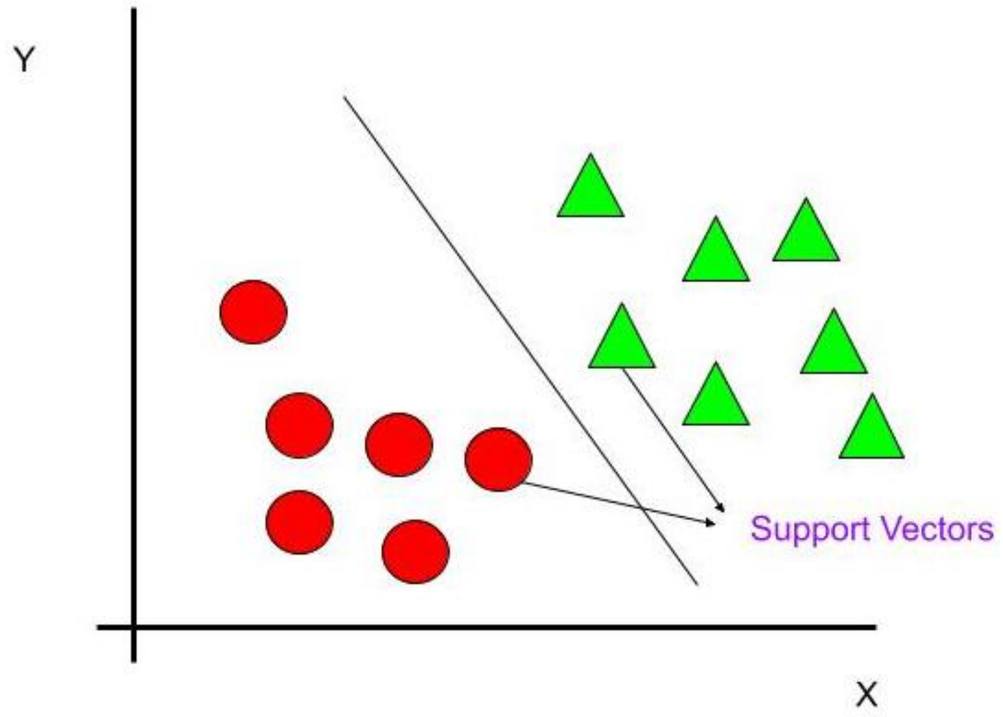
INVESTIGATION OF ONLINE SOCIAL NETWORKS.

Depression Analysis can be reviewed by the investigation of Online Social Networks for the forecasting of public health. They used the Twitter database to make predictions based on the tweets and status updates of the people, their social obligations, the timing they used, and the entire group's behavior.



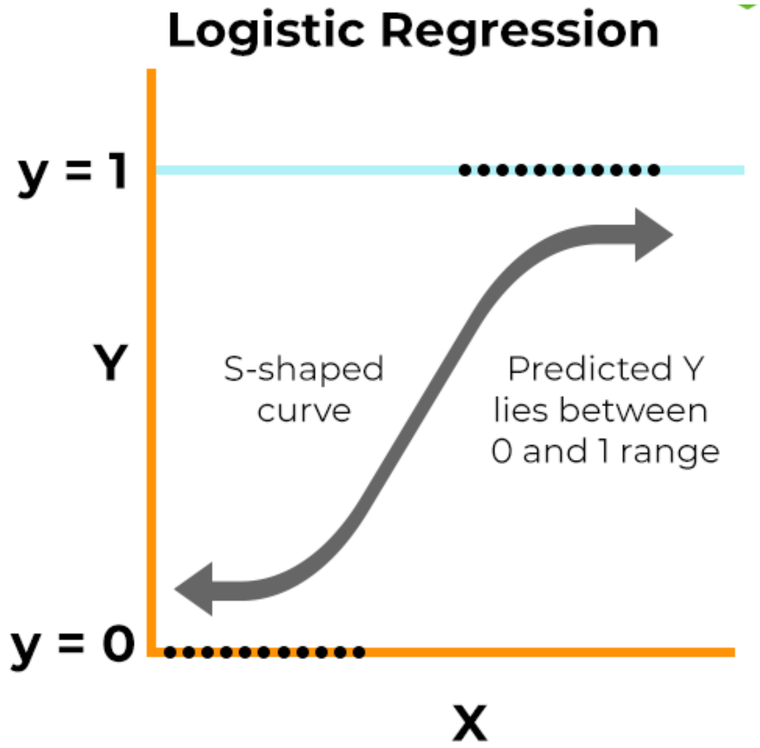
NAIVE BAYES

Naive Bayes classifier is **a probabilistic machine learning model based on Bayes' theorem**. It assumes independence between features and calculates the probability of a given input belonging to a particular class. It's widely used in text classification, spam filtering, and recommendation systems.



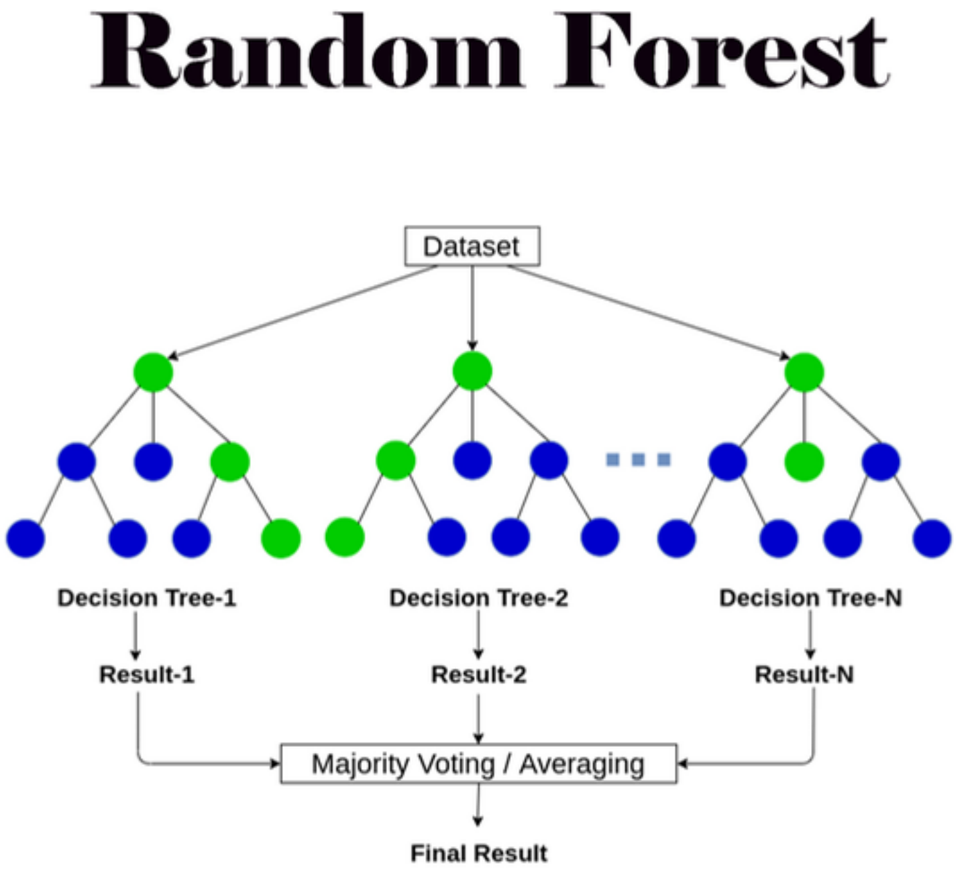
LINEAR SVM

Linear Support Vector Machine (SVM) is a powerful and widely used supervised learning algorithm for classification tasks. It is particularly well-suited for scenarios where the data is linearly separable, meaning that there exists a hyperplane that can cleanly separate the data points of different classes.



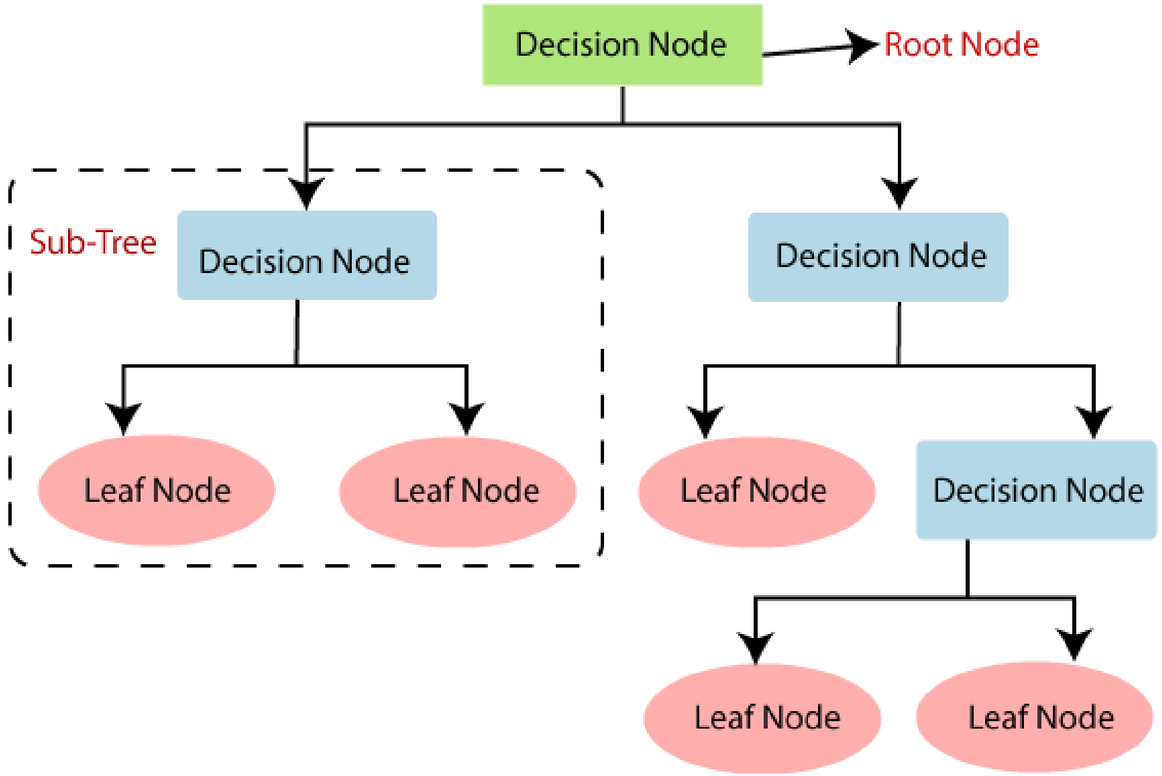
LOGISTIC REGRESSION

Logistic regression is **a process of modeling the probability of a discrete outcome given an input variable**. The most common logistic regression models a binary outcome; something that can take two values such as true/false, yes/no, and so on.



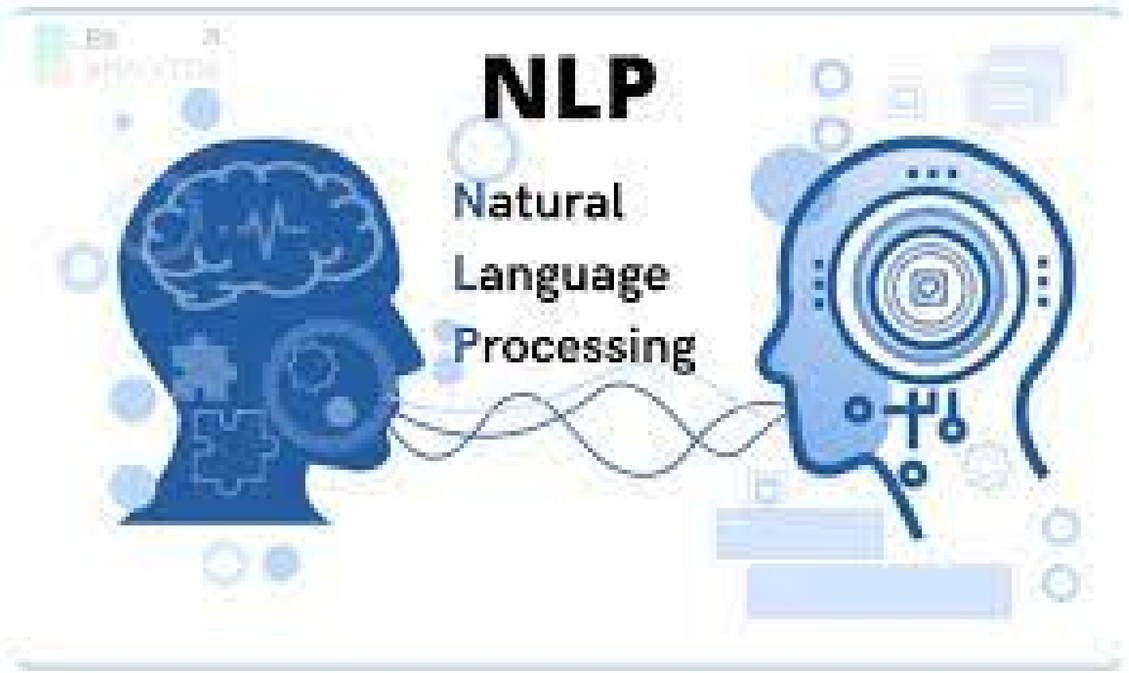
RANDOM FOREST

Random Forest is a versatile and powerful ensemble learning algorithm used for both classification and regression tasks. It operates by constructing a multitude of decision trees during training and outputs the class that is the mode of the classes (classification) or the mean prediction (regression) of the individual trees.



DECISION TREE

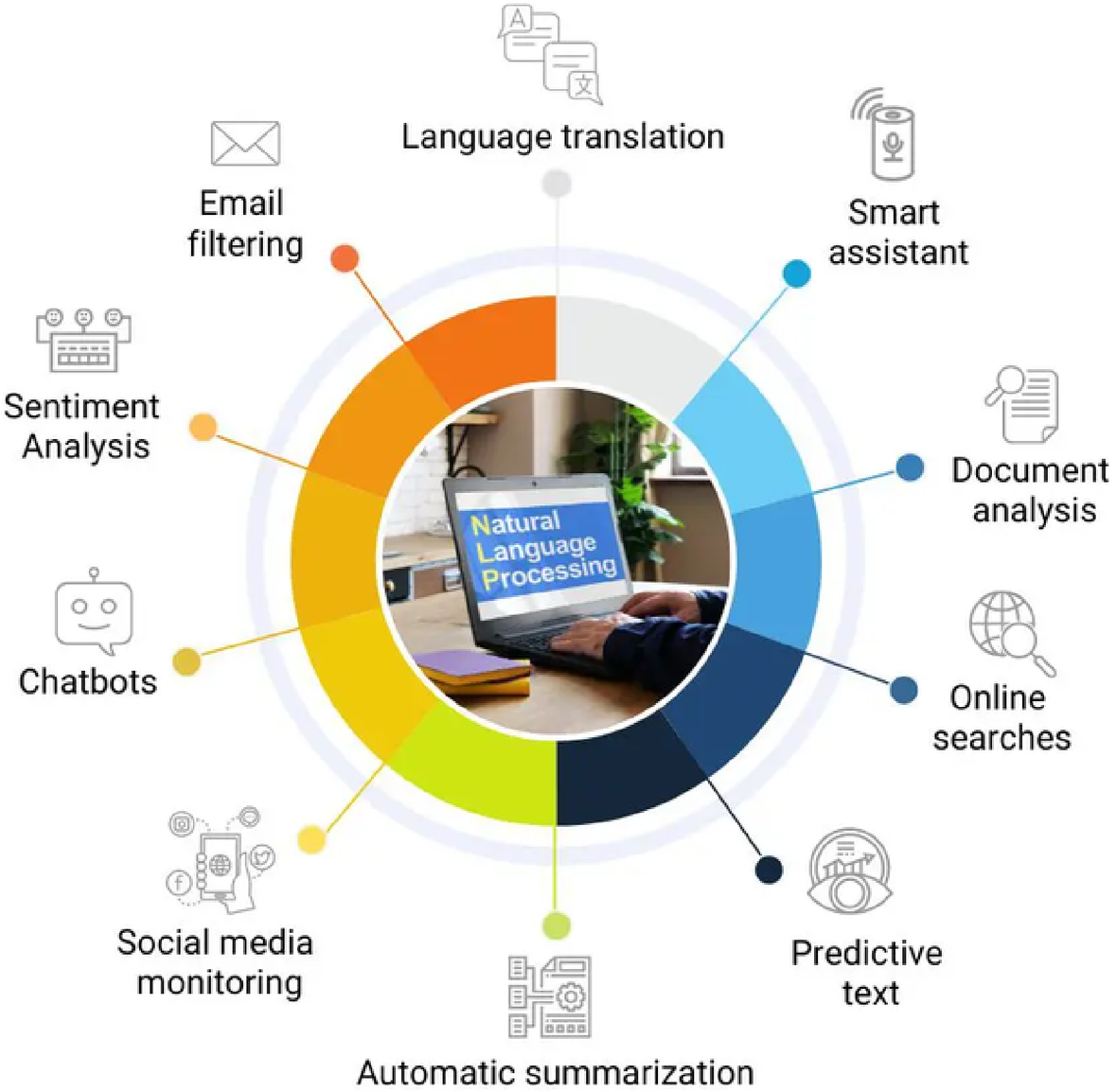
A Decision Tree is a versatile and intuitive supervised learning algorithm used for both classification and regression tasks. It models the relationships between input features and the target variable by recursively partitioning the feature space into regions, with each region corresponding to a particular class or predicted value.

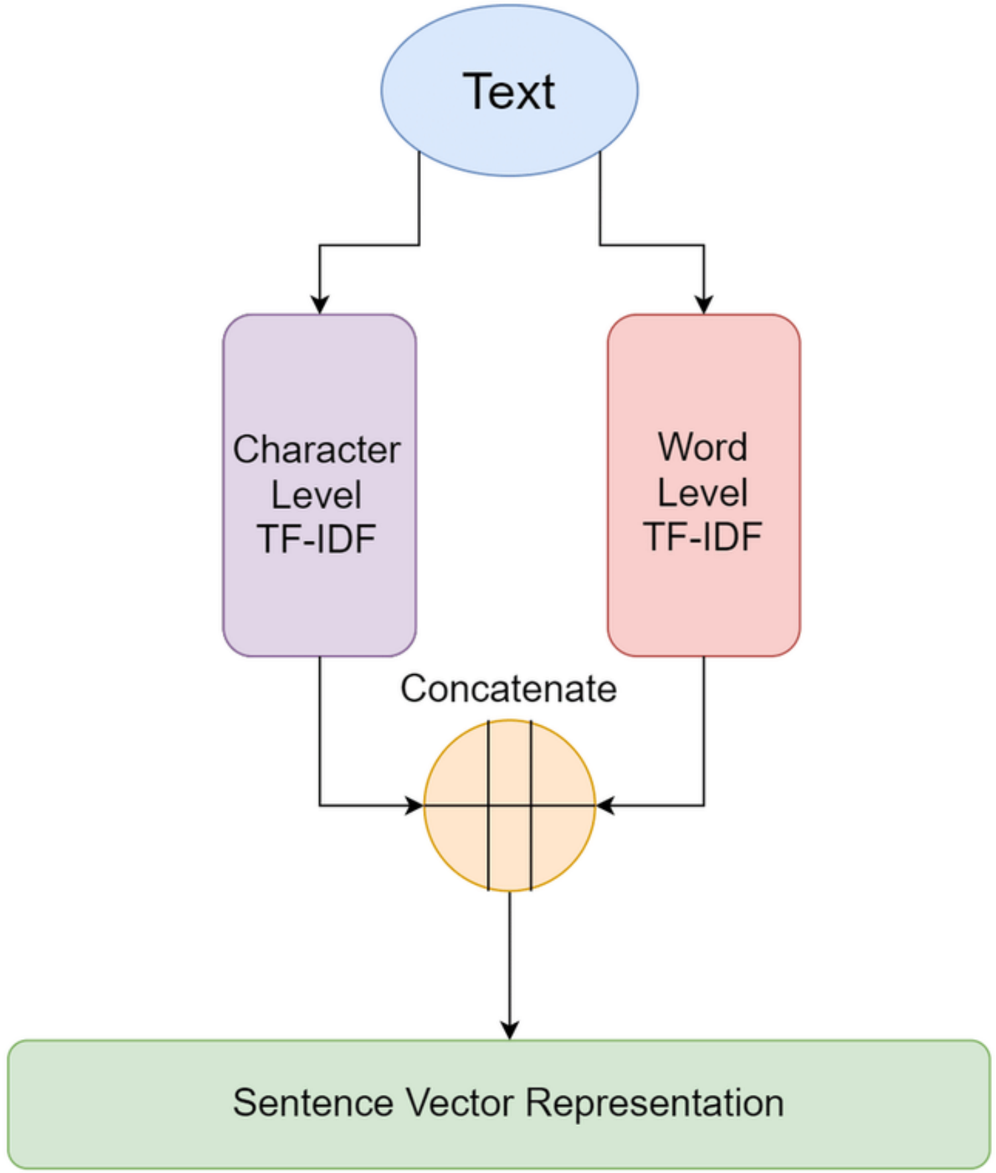


NATURAL LANGUAGE PROCESSING

Natural language processing, or NLP, combines computational linguistics—rule-based modeling of human language—with statistical and machine learning models to enable computers and digital devices to recognize, understand and generate text and speech.

Applications of
Natural Language Processing



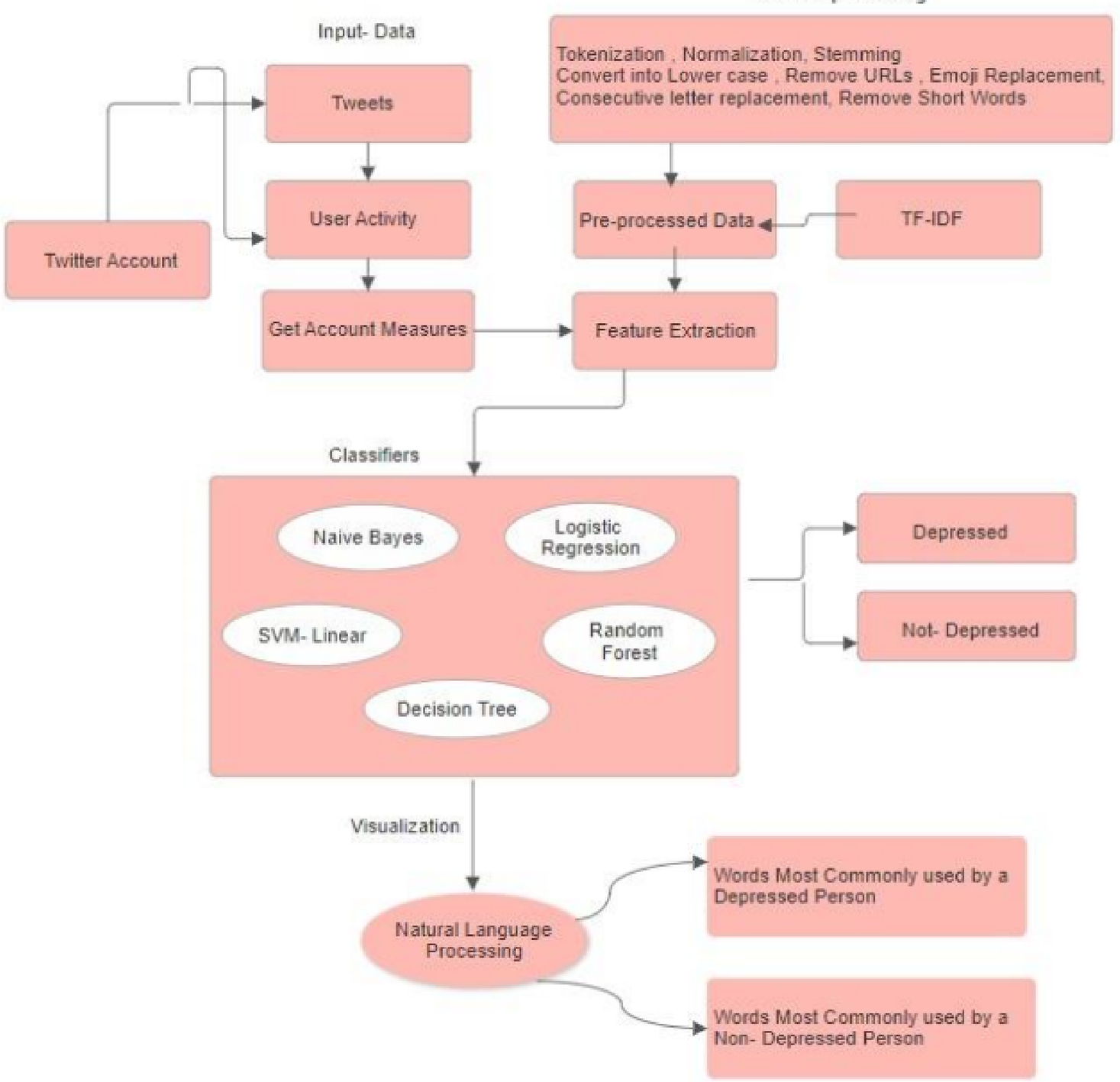


TF-IDF VECTORIZATION

TF-IDF vectorization is a conventional computer learning method, particularly text vectorization with the application of the TF-IDF (Term Frequency-Inverse Document Frequency) algorithm, a method frequently employed in jobs related to natural language processing (NLP).

METHODOLOGY

Depression Detection using Machine Learning Techniques



INPUT DATA

Sentiment140 dataset, with 1.6 million tweets, was the input data. It is a collection of tweets that have been compiled from the Twitter network. There are over 1.6 million tweets in it, all of which have sentiment polarity labels. The tweet’s sentiment polarity specifies whether it is neutral, positive, or negative. 1,600,000 tweets were retrieved from it using the Twitter API.

DATA PREPROCESSING

Step 1:Lower Case: All texts are changed to lowercase

Step 2:URLs that begin with "http", "https", or "www" are replaced with "URL".

Step 3: Emoji Replacement: Emojis can be replaced by utilising a pre-defined vocabulary that includes emojis and their meanings. (For example, ":)" to "EMOJIsmile")

DATA PREPROCESSING

Step 4: Usernames should be replaced with the word "USER" instead of @Usernames. (For example, "@Kaggle" to"USER")

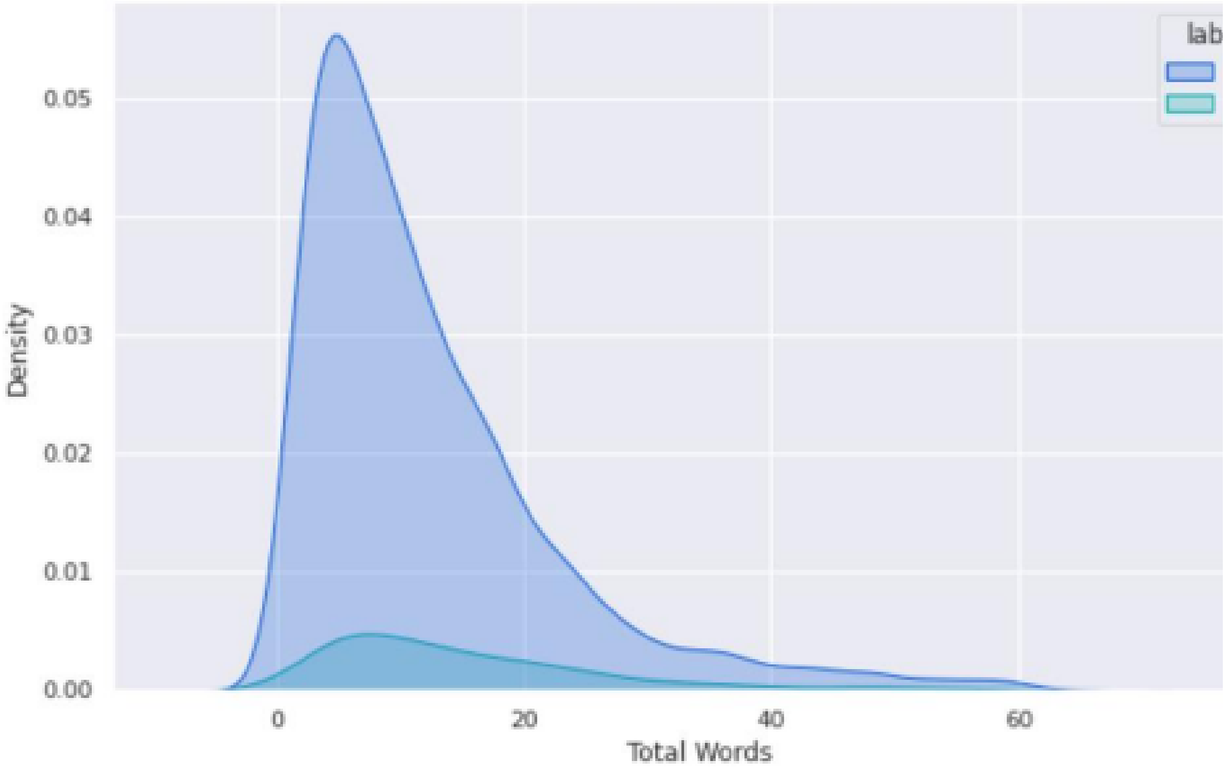
Step 5: Non-Alphabets are removed by replacing all characters except Digits and Alphabets with a space.

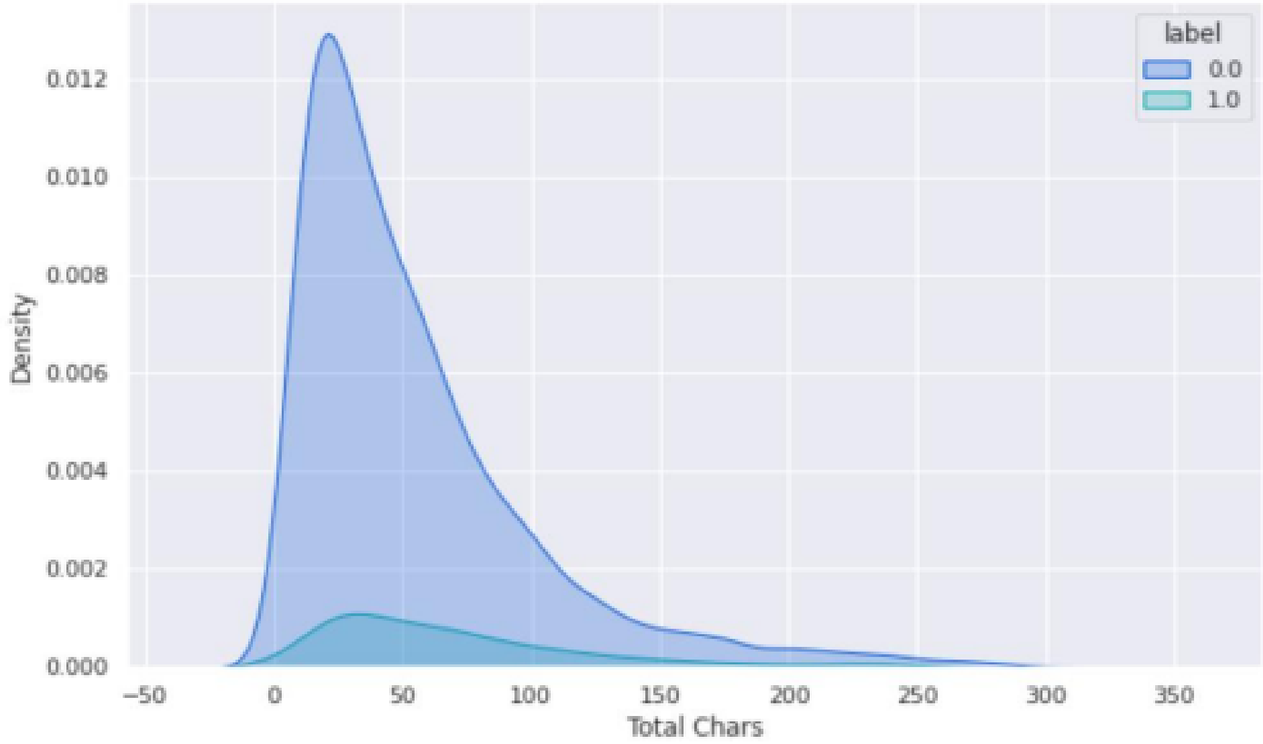
Step 6: Consecutive letter replacement: 3 or more consecutive letters are replaced by 2letters. (For example, "Hllll" to "Hllll) Step 7: Short Words are eliminated: Words with a length of fewer than two are eliminated.

FEATURE ENGINEERING

The act of choosing, modifying, and producing pertinent features or variables from raw data that are instructive for forecasting or diagnosing depression is known as feature engineering in the context of machine learning-based depression detection.

FEATURE ENGINEERING





Feature Engineering for Testing and Training Dataset

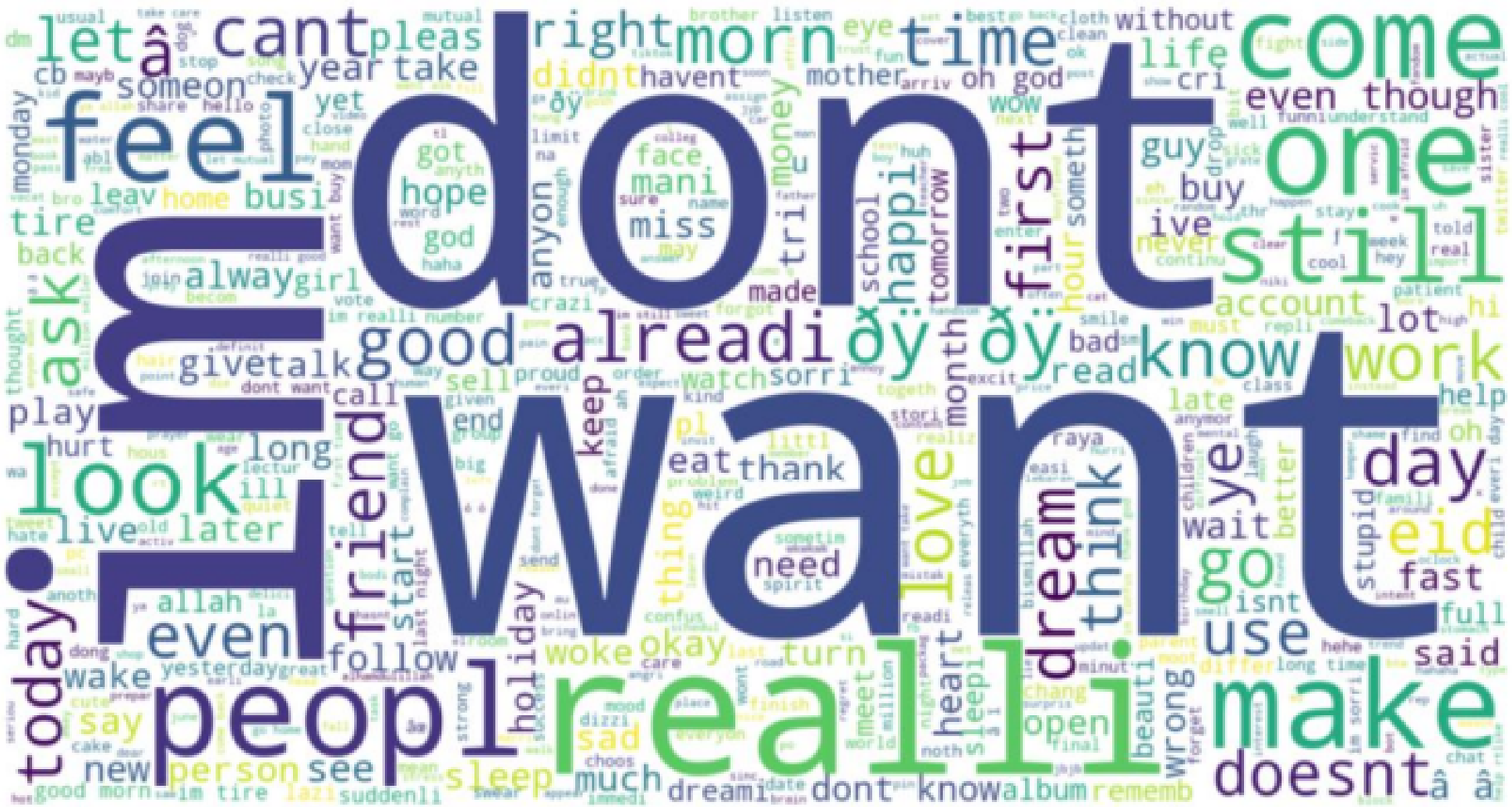
DATA VISUALIZATION

The goal of this project is to use natural language processing (NLP) techniques along with Matplotlib and Seaborn for data visualization to study and show the linguistic patterns of individuals who are depressed and those who are not.

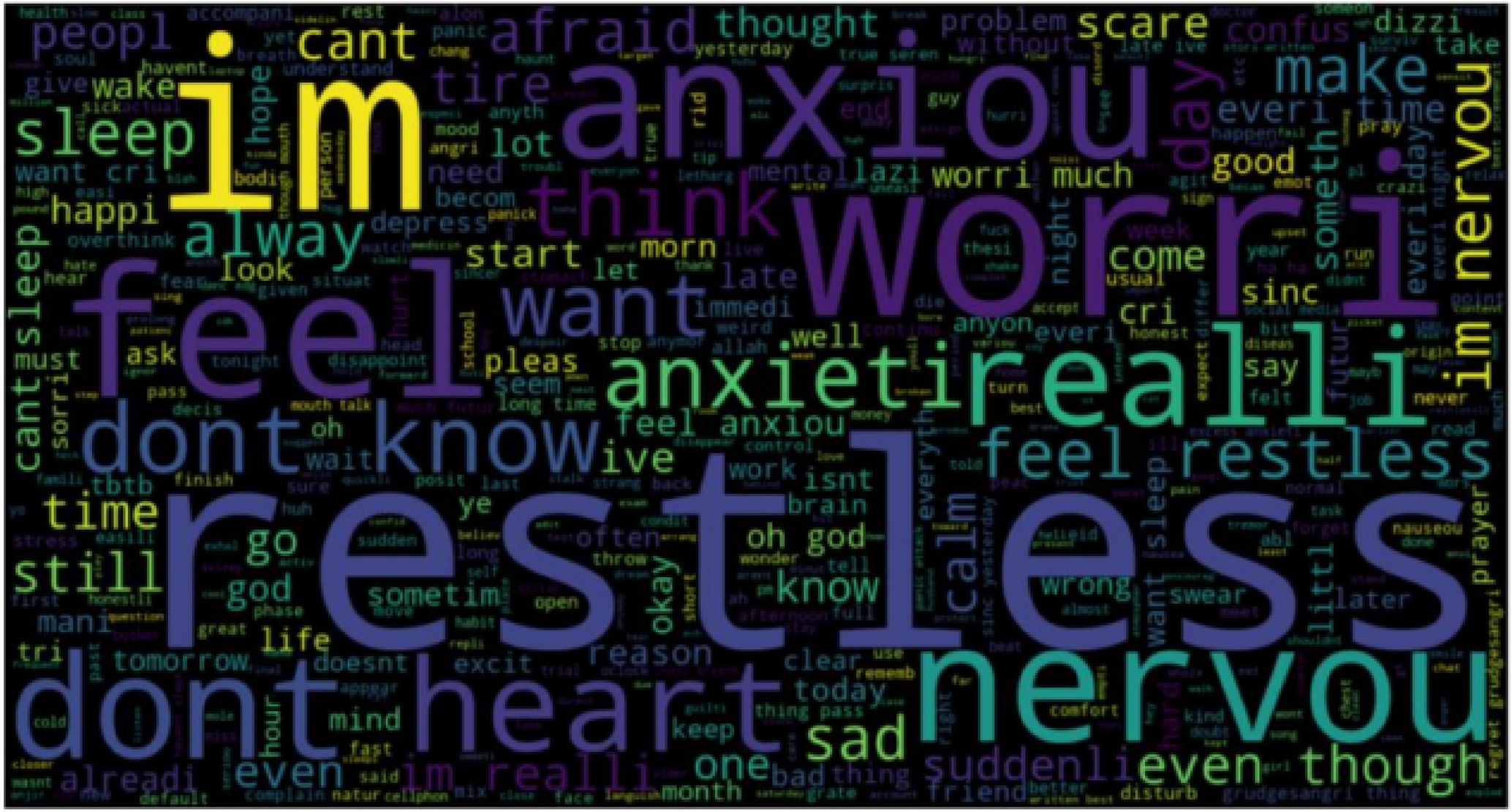
Both people without depression diagnoses and those with depression diagnoses provided text records for collection. After the data is Preprocessed, the Analysis of Word Frequency was done. For both depressed and non-depressed people, word clouds were created, with word size according to frequency

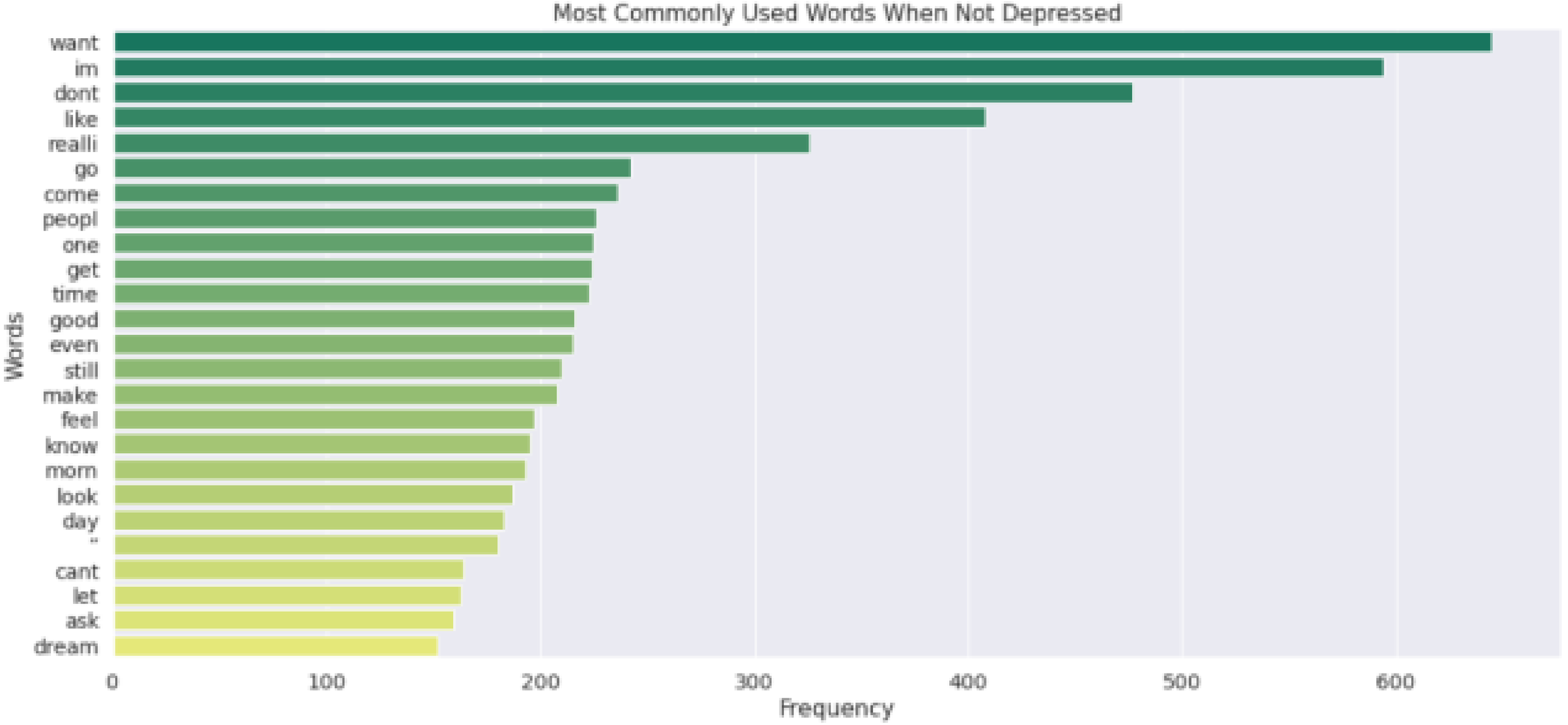
The subjects within each category were visualized using Stacked bar graphs.The subjects within each category were visualized using Stacked bar graphs.

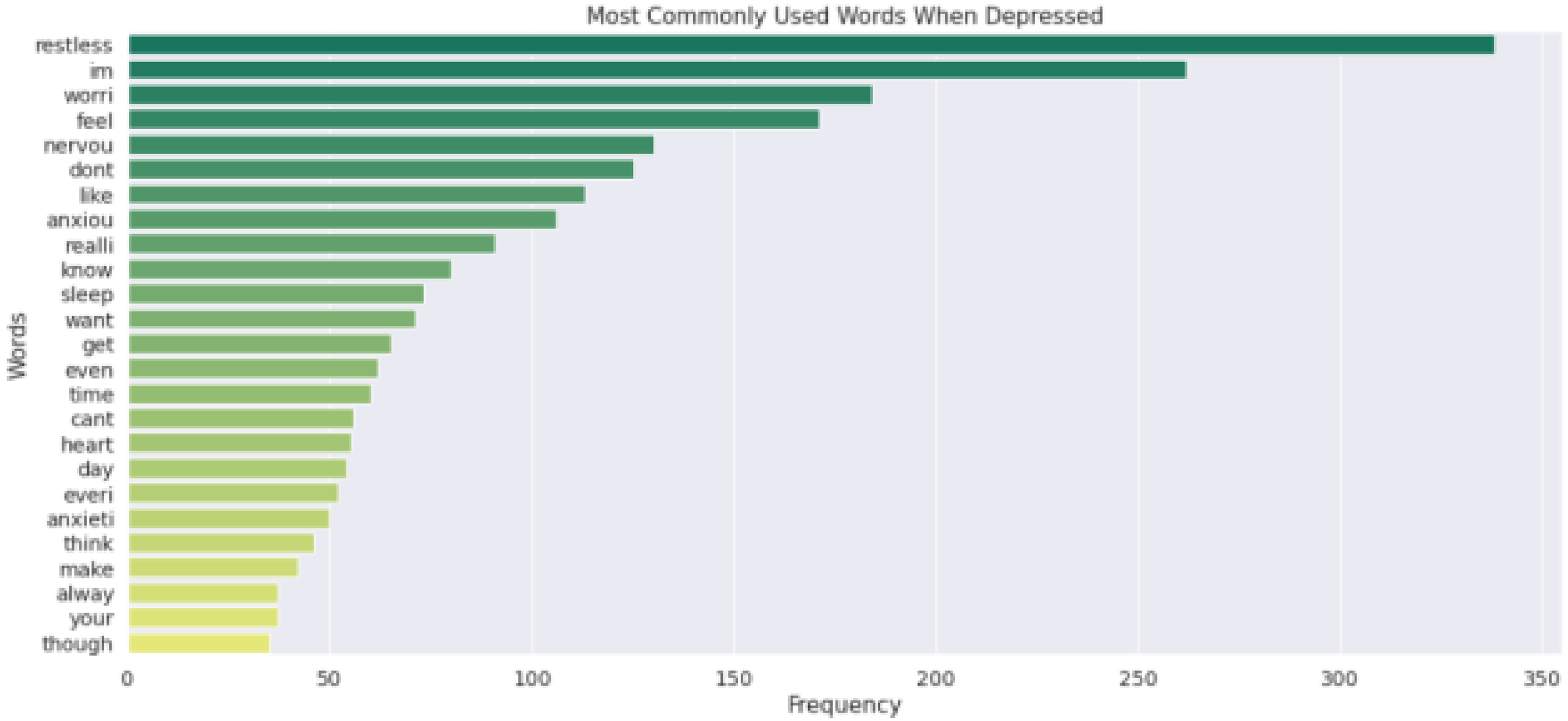
Word Cloud Visualization of Non- Depressive Words



Word Cloud Visualization of Depressive Words




Most Commonly Used Words When Not Depressed
want
im
dont
like
realli
go
come
peopl
one
get
time
good
even
still
make
feel
know
mom
look
day
"
cant
let
ask
dream
Words
0
100
200
300
400
500
600
Frequency


Most Commonly Used Words When Depressed
restless
im
worri
feel
nervou
dont
like
anxiou
realli
know
sleep
want
get
even
time
cant
heart
day
everi
anxieti
think
make
alway
your
though
Words
0
50
100
150
200
250
300
350
Frequency

IMPLEMENTATION

IMPLEMENTATION DETAILS

For lone classifiers, identifying depression and feeling emotions is an extremely challenging task. To achieve five classifications, however, our proposed technique combines the Linear SVM, Naive Bayes, logistic regression, Random Forest, and decision tree algorithm.

CONFUSION METRICS

Accuracy is a metric that measures how often a machine learning model correctly predicts the outcome. You can calculate accuracy by dividing the number of correct predictions by the total number of predictions.

Precision is defined as the proportion of correctly classified positives to all predicted positives

Recall is defined as the proportion of accurately identified positives to all positives.

The F1 Score, sometimes referred to as the F-measure it equally weights each metric as the harmonic mean of recall and precision

Table 1.1 Confusion Matix: Naïve Bayes

	Positive 0.0	Negative 1.0
Positive 0.0	1248	0
Negative 1.0	92	54

Table1.2 Confusion Matix: Random Forest

	Positive 0.0	Negative 1.0
Positive 0.0	1245	3
Negative 1.0	12	134

Table 1.3 Confusion Matix: Linear SVM

	Positive 0.0	Negative 1.0
Positive 0.0	1245	3
Negative 1.0	21	125

Table 1.4 Confusion Matix: Logistic Regression

	Positive 0.0	Negative 1.0
Positive 0.0	1247	1
Negative 1.0	49	97

Table 1.5 Confusion Matix: Decision Tree

	Positive 0.0	Negative 1.0
Positive 0.0	1245	3
Negative 1.0	9	137

Accuracy, Precision and Recall values of Machine Learning Classifiers

	Accuracy	Precision	Recall
Naïve Bayes	0.934	1.0	0.37
Random Forest	0.989	0.978	0.918
Linear SVM	0.983	0.977	0.856
Logistic Regression	0.964	0.99	0.664
Decision Tree	0.991	0.979	0.938

ROC CURVE

The Receiver Operating Characteristic curve, or ROC curve, is implemented to show how well a binary classification model performs across various thresholds. At different threshold values, it compares the True Positive Rate (TPR) against the False Positive Rate (FPR). The ROC curves have been plotted for Naïve Bayes, Linear SVM , Logistic Regression, Random Forest, and Decision Tree Algorithms.

RESULTS

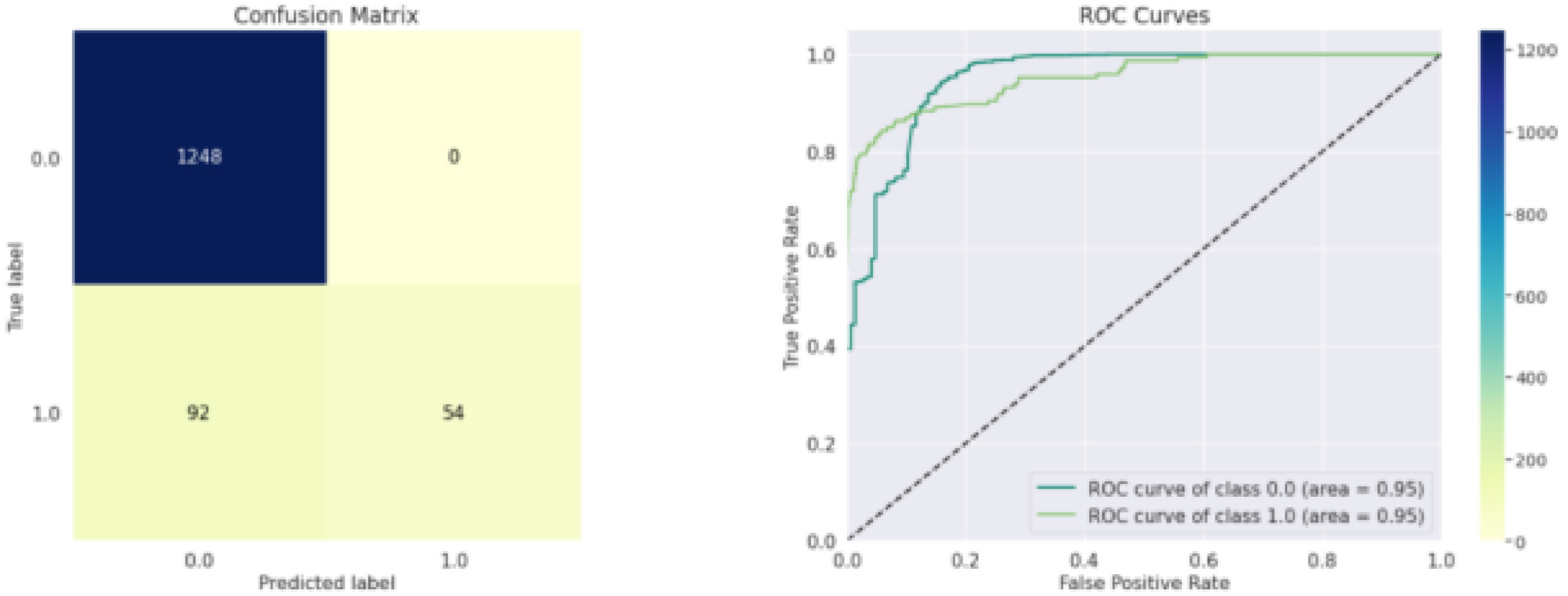
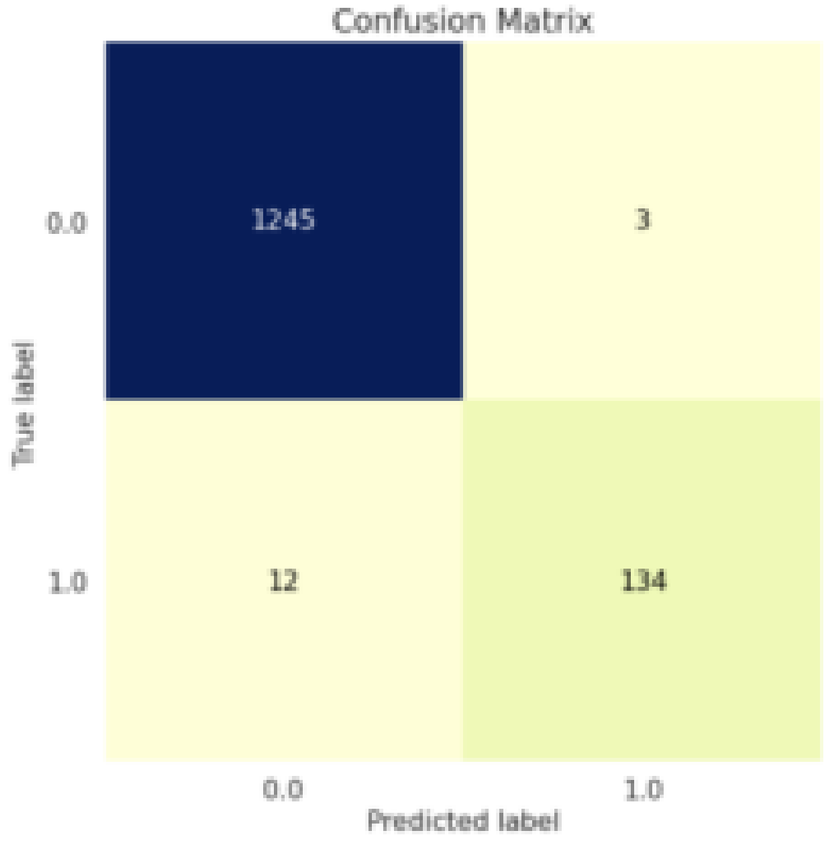


Figure 8. *Confusion Matrix and Roc Curve for Naïve Bayes*



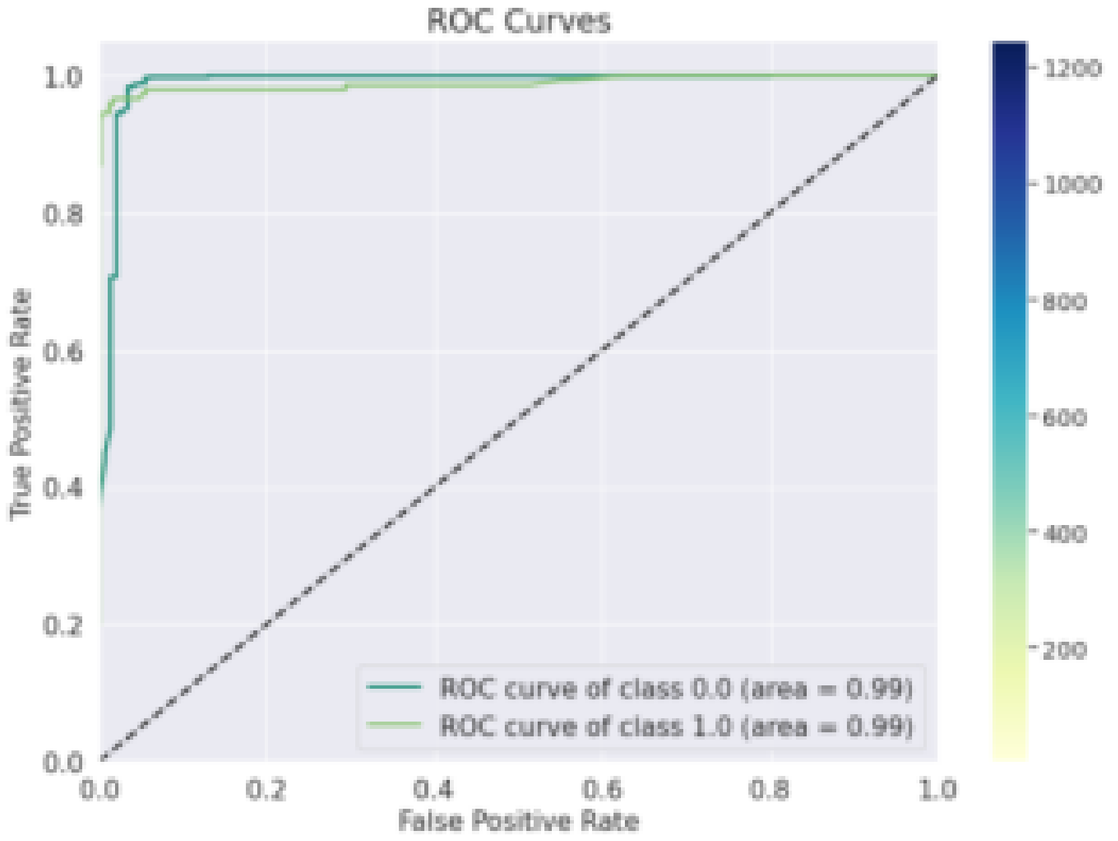


Figure 9. *Confusion Matrix and Roc Curve for Random Forest*

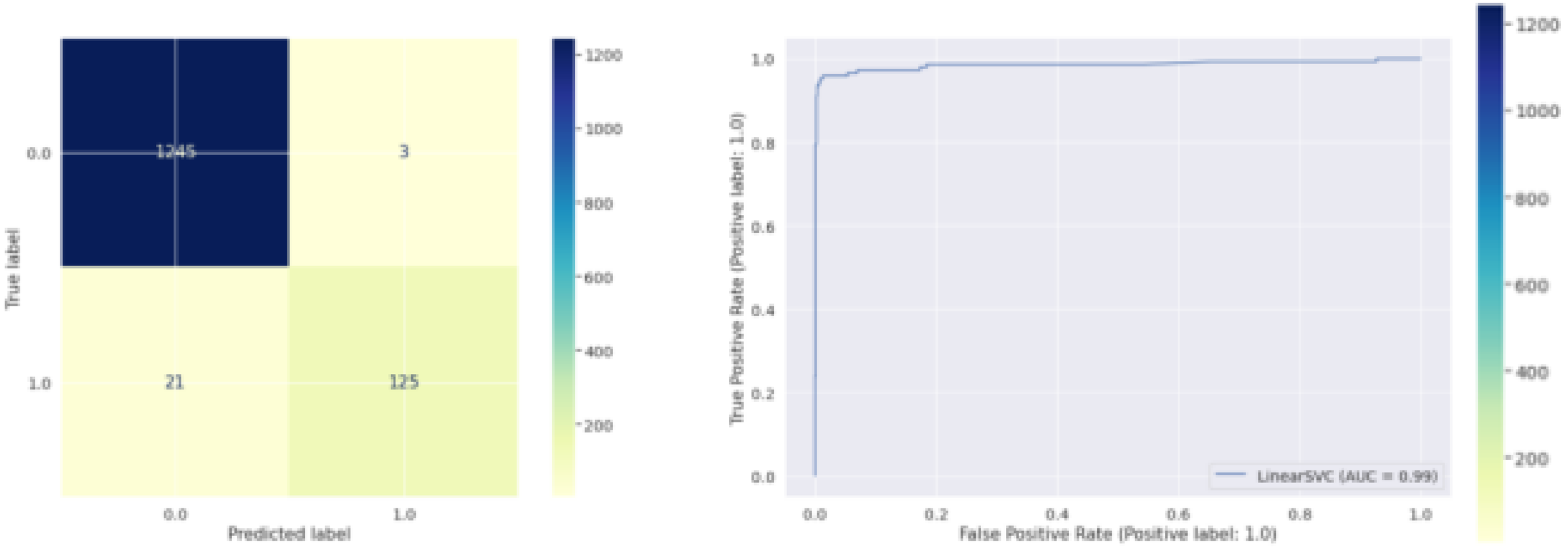


Figure 10. *Confusion Matrix and Roc Curve for Linear SVM*

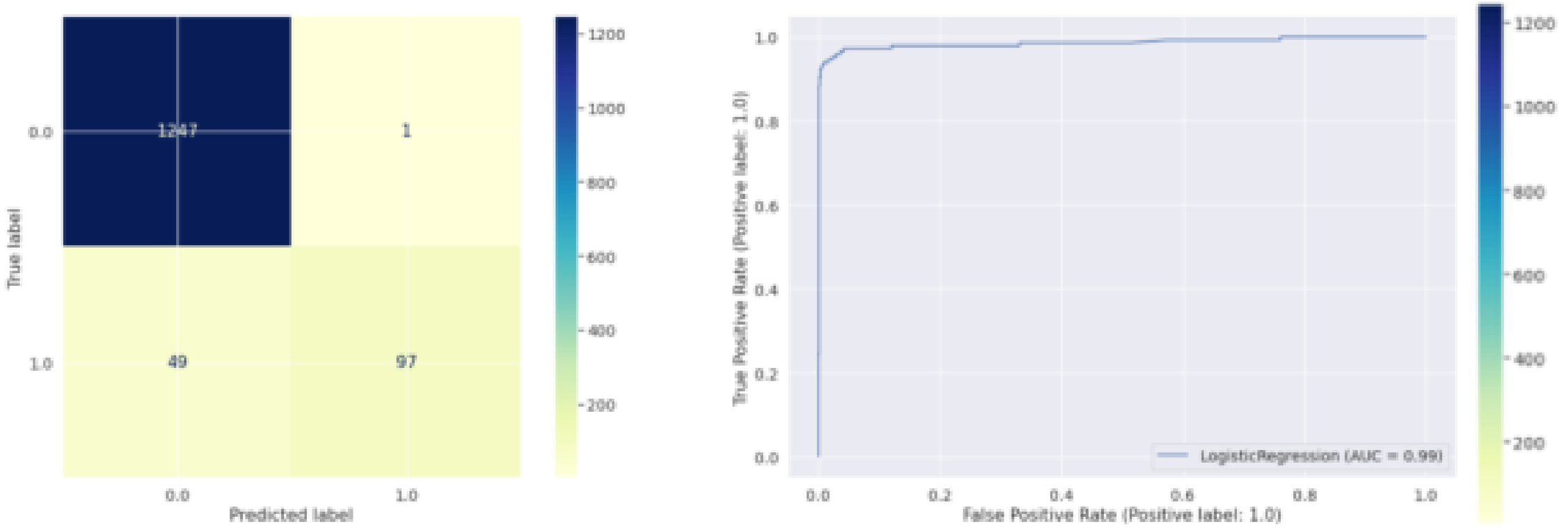


Figure 11. *Confusion Matrix and Roc Curve for Logistic Regression*

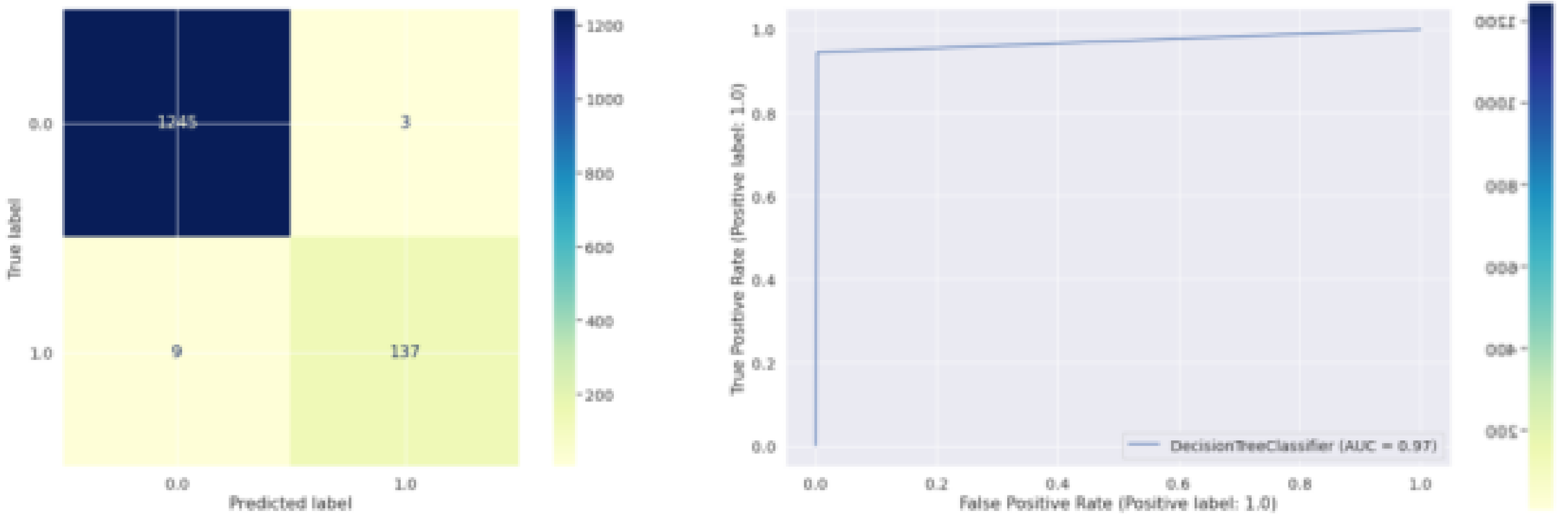


Figure 12. *Confusion Matrix and Roc Curve for Decision Tree*

CONCLUSION

In conclusion, our study investigated the effectiveness of various machine-learning algorithms for depression detection using a Twitter dataset. Through rigorous experimentation and analysis, we found that the Decision Tree algorithm consistently outperformed other methods, including Naive Bayes, Logistic Regression, Linear SVM, and Random Forest.

The Decision Tree model demonstrated superior performance in accuracy with 0.99 Accuracy Score distinguishing between individuals exhibiting signs of depression and those who were not, as evidenced by its higher classification accuracy, sensitivity, specificity, and precision. Moreover, the interpretability of Decision Trees allowed for better understanding and visualization of the underlying decision-making process, offering valuable insights into the features contributing to depression detection.

Additionally, our results indicate that the Random Forest algorithm emerged as the second-best performer among the models evaluated. While Decision Trees exhibited slightly better performance, Random Forest demonstrated robust classification capabilities and offered increased stability by aggregating multiple decision trees.

FUTURE WORKS

Our findings suggest that Decision Tree-based approaches, followed closely by Random Forest, hold promise for effective depression detection using social media data, providing valuable tools for mental health professionals and researchers In identifying individuals at risk and delivering timely interventions.

However, further research is warranted to explore additional feature engineering techniques, model optimizations, and the generalizability of results across diverse populations and social media platforms.