ABSTRACT

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, innovative research papers offer a promising avenue for exploration. The initial step in mitigating the impact of such conditions involves discovery. This study focuses on analyzing text-based depression using machine learning techniques by employing a Support Vector Machine-Linear SVM, Decision tree, Random Forest, Logistic Regression, and Naive Bayes algorithms on the Sentiment140 dataset with 1.6 million tweets. The study aims to identify the most commonly used words by a depressive person and by a Non-depressive person. Various parameters from cumulative distributions are incorporated into the classification model and dynamically categorized. The features utilized for identification are extracted from textual content for the Twitter dataset and semantic context. Comparative analysis reveals that the Random Forest detection method yields superior results compared to other machine learning approaches. The findings of this research may catalyze a re-imagining of how individuals' emotions are predicted, particularly as expressed on social media platforms like Twitter.

CONTENTS

1. INTRODUCTION

1.1 INTRODUCTION

Due to the fluctuating nature of mental states, it can be difficult to distinguish non-depressed and depressed users on online social networking platforms. Twitter has emerged as one of the most popular social networking platforms, with millions of users sharing updates on their daily lives, preferences, interests, and opinions regularly. The abundance of public viewpoints shared on Twitter offers valuable insights, but it can be challenging to filter and access relevant information in real time. Therefore, retrieving data from Twitter and conducting sentiment analysis is vital for understanding the overall sentiment of users. Despite its benefits, analyzing social media sentiment presents its own set of challenges.

1.2 Objective:

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.

1.3 Overview of the Project

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.

1.4 Chapter wise Summary

In Chapter two, the related works are reviewed. In Chapter three the method is described in details and the experimental results are showed in Chapter four. Chapter five contains the analysisand discussion followed by Chapter six with the conclusion.

2. RELATED WORK

2.1 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. In a different study, they used Facebook data to investigate emotions and develop a statistical model to forecast emotions.

2.2 Linguistic characteristics.

Depression can be predicted using linguistic characteristics. For the training decision list, they utilized the Twitter dataset to determine the depression. The method to find social media classifiers for post-traumatic stress disorder is presented. They displayed their research using Twitter users.

2.3 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.

2.4 Statistical Techniques.

Depression can be looked at by a number of statistical techniques using a machine learning algorithm to accurately identify a person's depression condition. This uses the Twitter dataset to analyze the variation in prediction rate. Each tweet's ability to persuade people is crucial in pitiful situations of any kind.

2.5 Sentimental Analysis Techniques.

The analysis of the Twitter data can be done focused on identifying emotions caused by psychological disorders or mental health issues. Using sentimental analysis techniques, they categorise the emotional status based on the behavioural characteristics of the subjects. Two tests are employed in [7] and the supervised machine learning classifiers' are used to explore emotional interaction. They classified depression utilising classification techniques.- relevant social media posts

3. IMPLEMENTATION.

3.1. Modules Description

Naive Bayes (NB):

The foundation for the Gaussian Naive Bayes (GNB) model of supervised learning is provided by the Bayes theorem. In this model, each feature's distribution is assumed to follow a Gaussian distribution and is considered independent of the other features in a data point, according to GNB. While the GNB model is quick and simple to apply, the presence of non-Gaussian

$$P(A|B) = \frac{P\left(\frac{B}{A}\right) * P(A)}{P(B)} \longrightarrow [1]$$

Despite being considered one of the simplest techniques in machine learning, the NB classifier remains competitive with SVM.

Logistic Regression(LR):

When attempting to forecast the likelihood that an instance would belong to a specific class, one statistical model called logistic regression is employed. Logistic regression forecasts a binary outcome's probability based on one or more predictor variables, as opposed to linear regression's prediction of continuous outcomes. The predicted probabilities are guaranteed to lie between 0 and 1 by the Sigmoid function.

$$h\theta(X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}} \longrightarrow [2]$$

Because of its ease of use, interpretability, and efficiency in simulating binary outcomes, logistic regression is extensively employed in a variety of domains, including healthcare, finance, marketing, and social science.

Decision Tree (DT):

The Decision Tree (DT) is a supervised learning model that creates predictions using a topology that resembles a tree. The data is divided into smaller and smaller subgroups until all of the data points in each subset are exclusive to one class. This process creates the tree. Although DTs are frequently easy to comprehend and analyze, data noise may affect them. Instances are sorted by DTs according to the feature values. Each division of a DT represents a value that the node may perform, and each node represents a feature.

A DT uses split selection, a crucial component that seeks to identify an attribute and associated splitting function for each test node in a DT, to divide the data into subdivisions that contain occurrences with comparable values.

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2 \longrightarrow [3]$$

To assess splits, one can compute entropy. Complicated DTs are widely used in the machine learning industry since they merely ask a sequence of well-crafted questions to categorize tasks that are simple in nature.

Random Forest(RF) :

An ensemble learning technique called Random Forest is applied to both regression and classification problems. During training, it creates a large number of decision trees, from which it produces the mode (classification) or average prediction (regression) of each tree. During training, it builds a large number of decision trees, from which it produces the mode (classification) or average prediction (regression) of each tree. In order to produce a prediction that is more reliable and accurate, Random Forest constructs several decision trees and combines them. A random subset of the training data and a random subset of the characteristics are used to build each decision tree in the Random Forest.

$$Gidex = 1 - \sum_{i=1}^{n} (p_i)^2 \qquad = 1 - [(P+)^2 + (P-)^2 \longrightarrow [4]$$

Because of its scalability, resilience, and capacity to manage high-dimensional data with intricate feature interactions, Random Forest is frequently utilized pical applications. It is a well-liked option for many machine-learning problems, suchas feature importance esimation, regression, and classification.

Linear Support Vector Machine (SVM):

The SVM-supervised learning model identifies two unique classes in a high-dimensional 5-space. It can balance exceptional performance with changes to several features to reduce the probability of overfitting. Strong theoretical underpinnings and insensitivity to high-dimensional data are two of SVM's well-known strengths, especially when used with real-world data.

The linear classifier uses the inner product of the vectors, which are the support vector and the test pair. An inner product is a kernel function in some extended feature space. A common kernel function is the radial basis function in infinite dimensional space.

$$k(a_i, a_j) = {a_i}^T a_j \longrightarrow [5]$$

Natural Language Processing (NLP):

Computers can now analyze, alter, and comprehend human language thanks to a machine learning technique called natural language processing, or NLP. To identify depression and its severity, natural language processing (NLP) techniques can be used with machine learning techniques. NLP techniques concentrate on the analysis of linguistic and acoustic elements of human language derived from speech and text.

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).

A traditional machine-learning method for preprocessing, extracting features, and representing text input is called TF-IDF vectorization. It is frequently used for tasks like text categorization, sentiment analysis, and information retrieval in conjunction with other machine learning techniques like logistic regression, support vector machines, and naive Bayes classifiers.

The process of text vectorization involves transforming the training and test data for a machine learning model and applying the TF-IDF (Term Frequency-Inverse Document Frequency) approach.

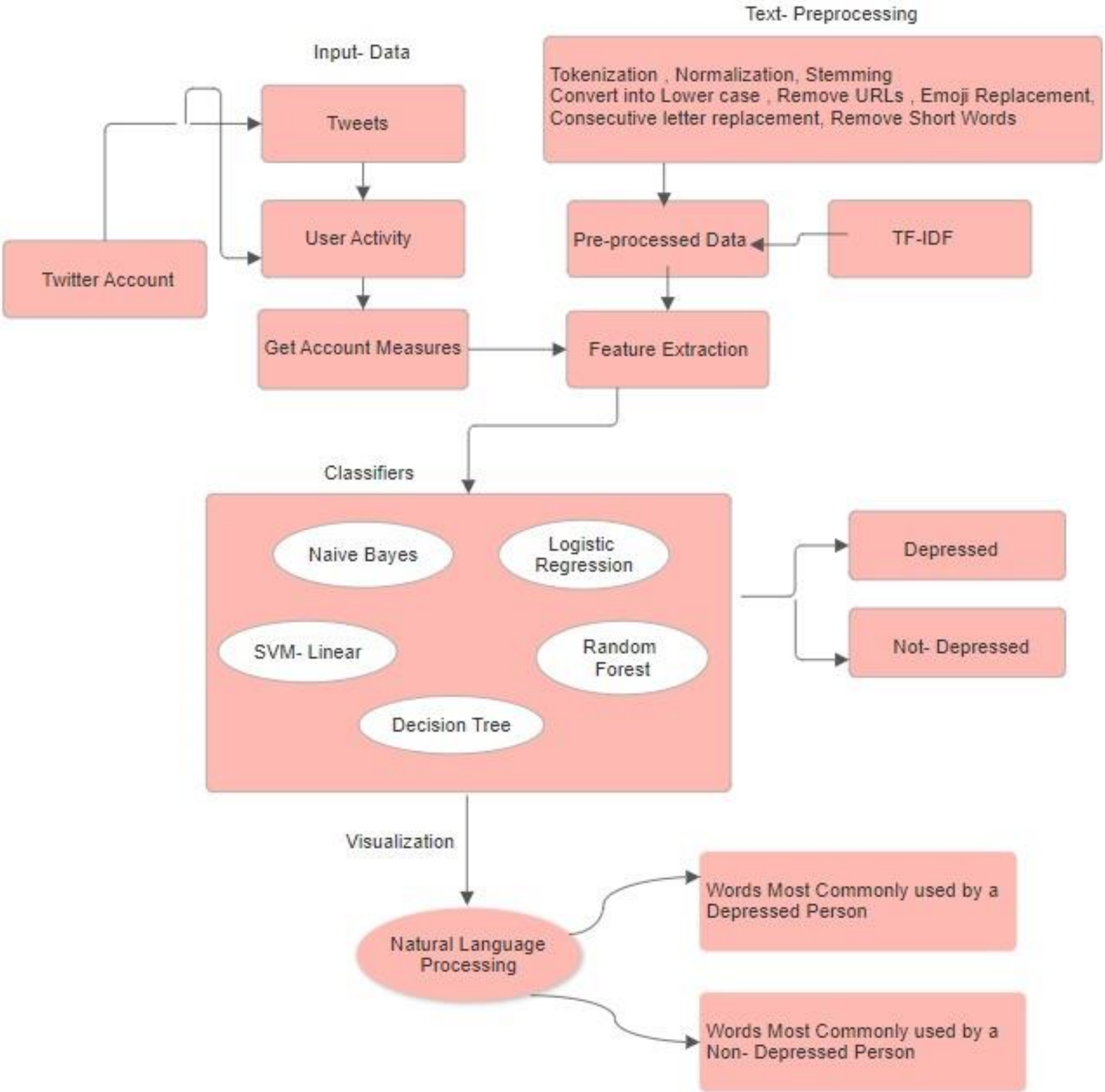

Figure 1. Methodology of Depression Detection

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. The tweets can be used to gauge sentiment because they have been annotated (0 being bad, 4 being favorable).

The following six fields are present in it:

1) **Target** : The tweet's polarity (0 being negative, 2 being neutral, and 4 being favorable).
2) **Ids** : The tweet's id (2087)
3) **Date** : The tweet was sent on Saturday, May 16, 2009, at 23:58:44 UTC.
4) **Flag** : The lyx query. This value is NO QUERY if there isn't a query.
5) **User** : @robotickilldozr, who tweeted.
6) **Text** : The tweet's content (Lyx is awesome).

Data Pre-Processing:

Pre-processing of the data is done in the first step of the detection model. This covers the transformation and normalization of data.

To clean up the dataset, stop words, retweets, URLs, and mentions are removed. After that, the text is broken up into words or tokens for every dataset row.

After tokenization, the words undergo stemming and lemmatization. In a machine learning prediction model, the traits are binary patterns that can be utilized to represent depression.

The preprocessed steps are as follows:

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")

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, "HIIII" to "HIIII)

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 can be used to extract features for depression detection. These characteristics could be used to distinguish between those who have depression and those who don’t, as well as to capture other facets of a person’s mental health status.

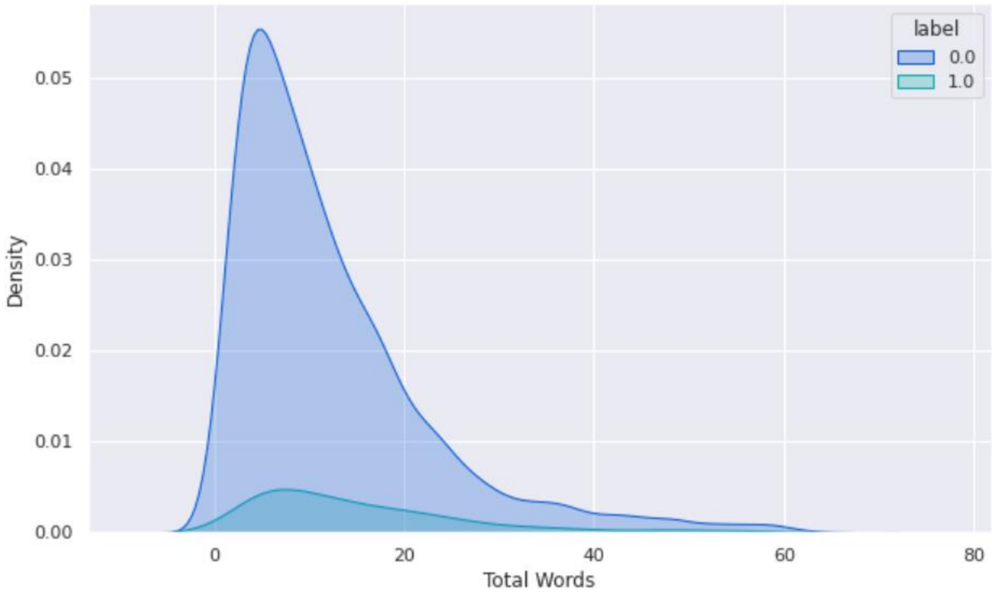

Figure 2. Feature Engineering of Total Words and Density

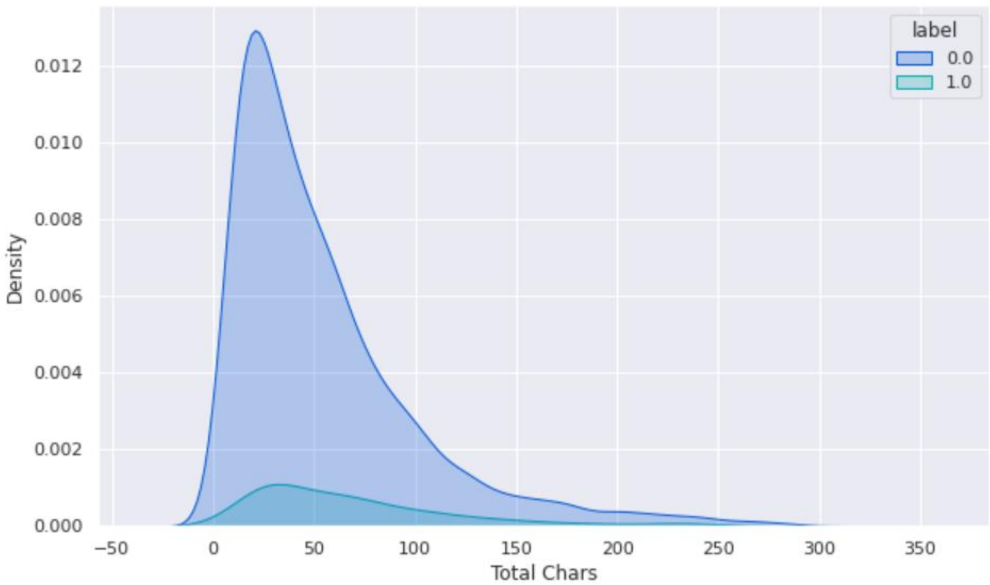

Figure 3. Feature Engineering of Total Chars and Density

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.

Our goal is to learn more about the linguistic traits linked to depression by looking at the terms that people with and without depression use most frequently.

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.

The frequency distributions of terms used by people with depression and those without were computed. To see which terms were used the most in each group, bar graphs and histograms were made.

For both depressed and non-depressed people, word clouds were created, with word size according to frequency.

To emphasize common words in each group, visually appealing word clouds were created using Matplotlib and Seaborn.

To examine the frequency of particular terms in depressed and non-depressed persons, bar plots were developed. For visual aids, words were categorized (e.g., good, negative feelings).

To categorize words as neutral, negative, or positive, sentiment analysis was used.

Pie charts and histograms were utilized to show how each group's sentiment was distributed.

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

The frequency of words in records gathered at various times was analyzed to investigate temporal trends in word usage, and **Word Cloud** was Generated

The time trends for depressed and non-depressed people were displayed using Matplotlib and seaborn.

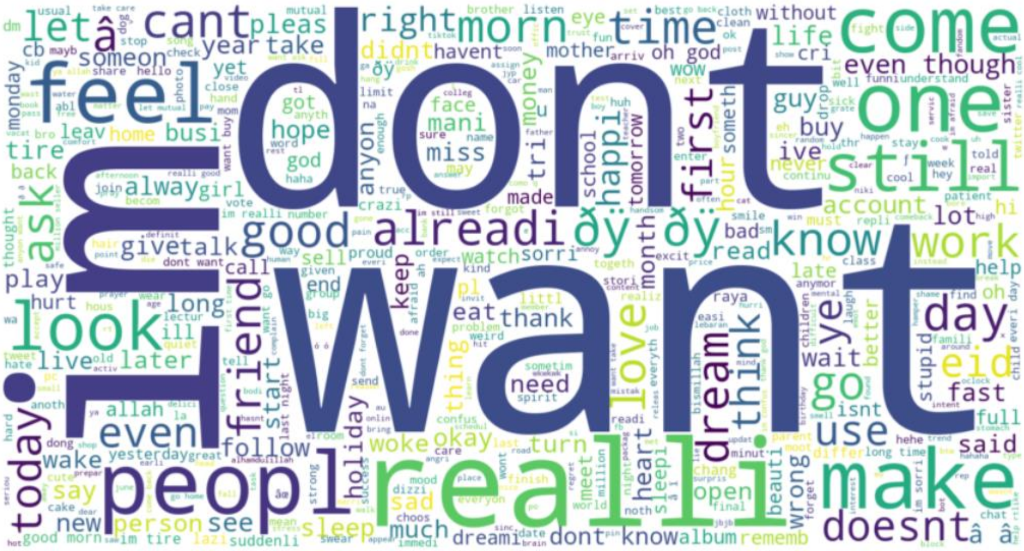

Figure 4. Word Cloud Visualization of Non- Depressive Words

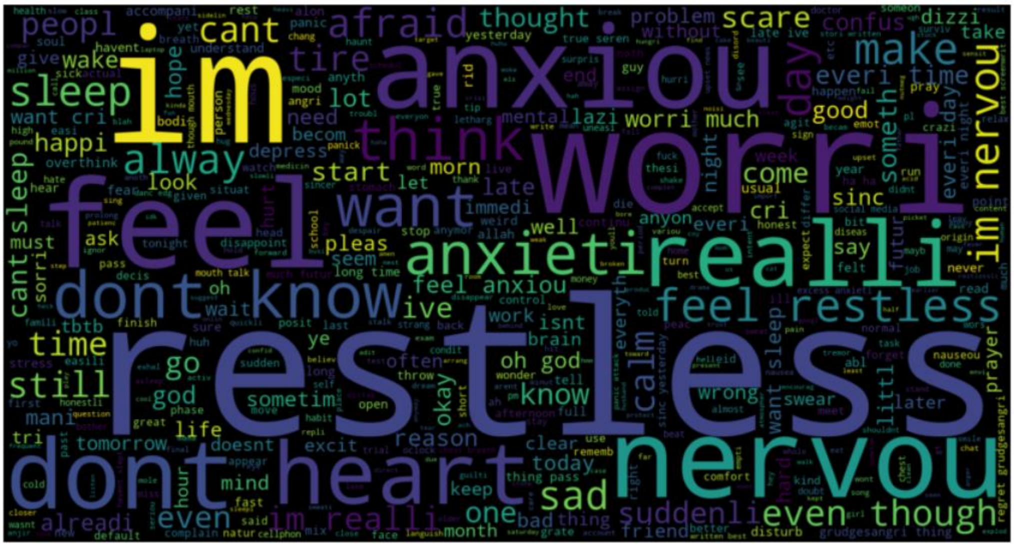

Figure 5. Word Cloud Visualization of Depressive Words

A traditional machine-learning method for preprocessing, extracting features, and representing text input is called TF-IDF vectorization. It is frequently used for tasks like text categorization, sentiment analysis, and information retrieval in conjunction with other machine learning techniques like logistic regression, support vector machines, and naive Bayes classifiers.

The process of text vectorization involves transforming the training and test data for a machine-learning model and applying the TF-IDF (Term Frequency-Inverse Document Frequency) approach. The Most Commonly used Words used by a Person when Not Depressed and when Depressed has been predicted and Visualized using Matplotlib and Seaborn.

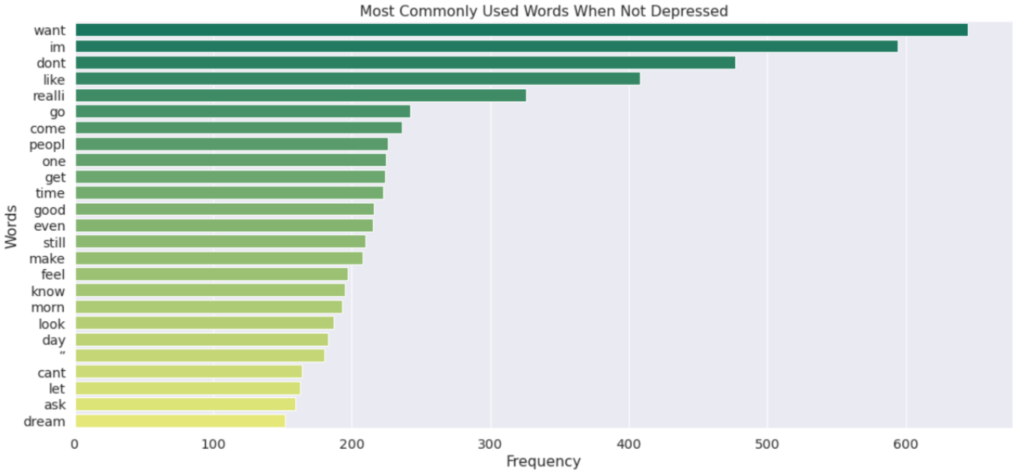

Figure 6. Most Commonly used Words When Not- Depressed

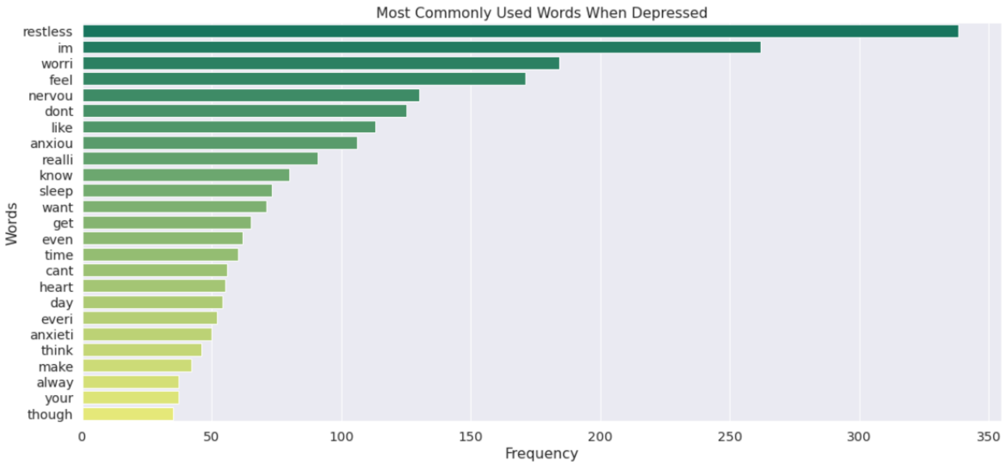

Figure 7. Most Commonly used Words When Depressed

3.2 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.

4. TESTS/RESULTS/VERIFICATION.

4.1 Testing and Training

The suggested technique incorporates additional variation elements to improve the performance metrics. In an online social media platform called Twitter, the following sample emotion keys are secured utilizing a confusion matrix for our prediction. observing a loss during the dataset's testing and training.

4.1.1 Confusion Metrics

Utilizing the confusion matrix (CM), standard accuracy (Acc), precision (P), recall (R), and F1 scores, the viability of automated prediction is examined and cross-validated.

i) **Accuracy:**

Accuracy is the most fundamental and often used parameter for evaluating a classifier. Its definition is the percentage of misclassification errors or, on the other hand, the degree to which a model produces accurate predictions.

$$Acc = (true\ positive + true\ negative) / (true\ positive + true\ negative + false\ positive + false\ negative)$$

ii) **Precision**

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

$$P = true\ positives / (true\ postives + false\ postive)$$

iii) **Recall**

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

$$R = true\ positives / (true\ positives + false\ negative)$$

iv) **F1 Score**

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

$$F1 = (2 * P * R) / P + R$$

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

Table 1.6 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

4.1.2 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.

4.2 RESULTS

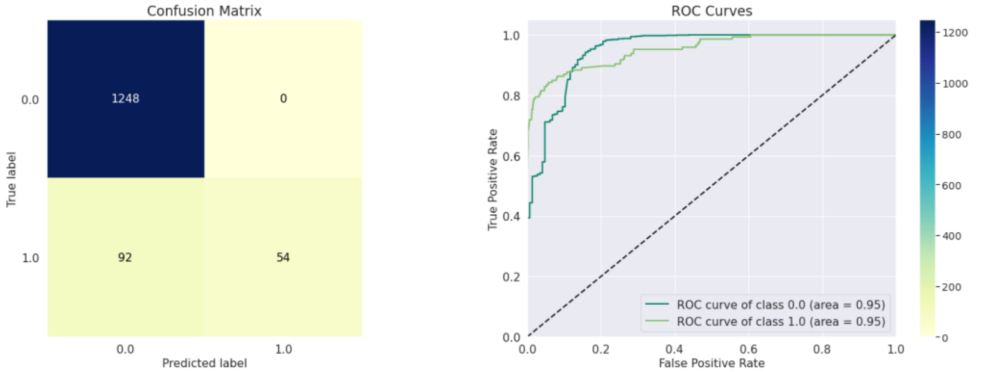

Figure 8. *Confusion Matrix and Roc Curve for Naïve Bayes*

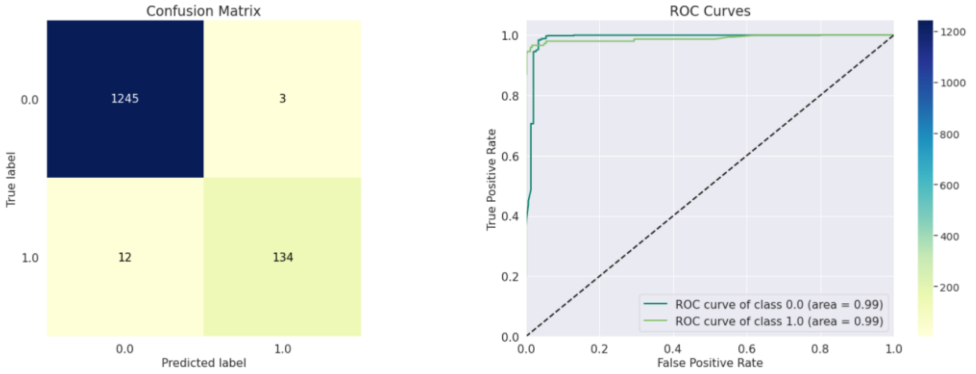

Figure 9. *Confusion Matrix and Roc Curve for Random Forest*

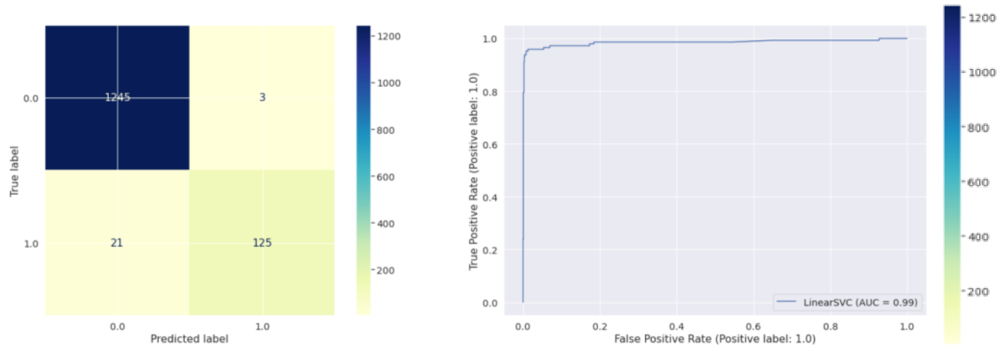

Figure 10. *Confusion Matrix and Roc Curve for Linear SVM*

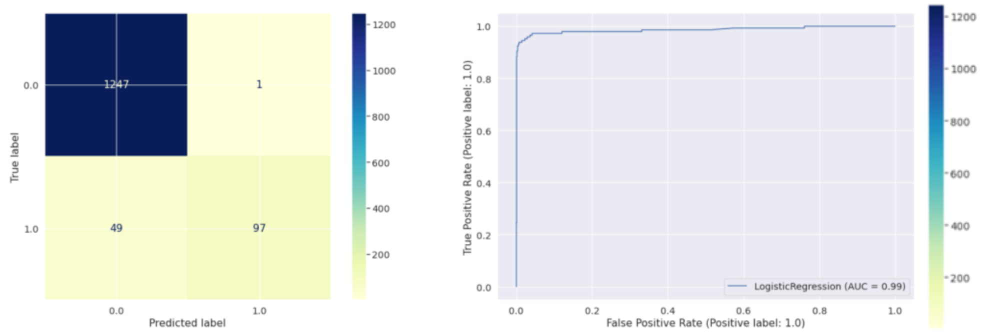

Figure 11. *Confusion Matrix and Roc Curve for Logistic Regression*

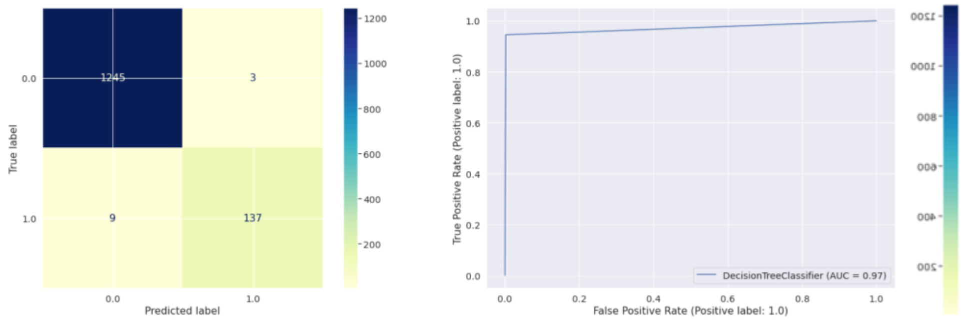

Figure 12. *Confusion Matrix and Roc Curve for Decision Tree*

5. CONCLUSION.

Performance of the Classifier:

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.

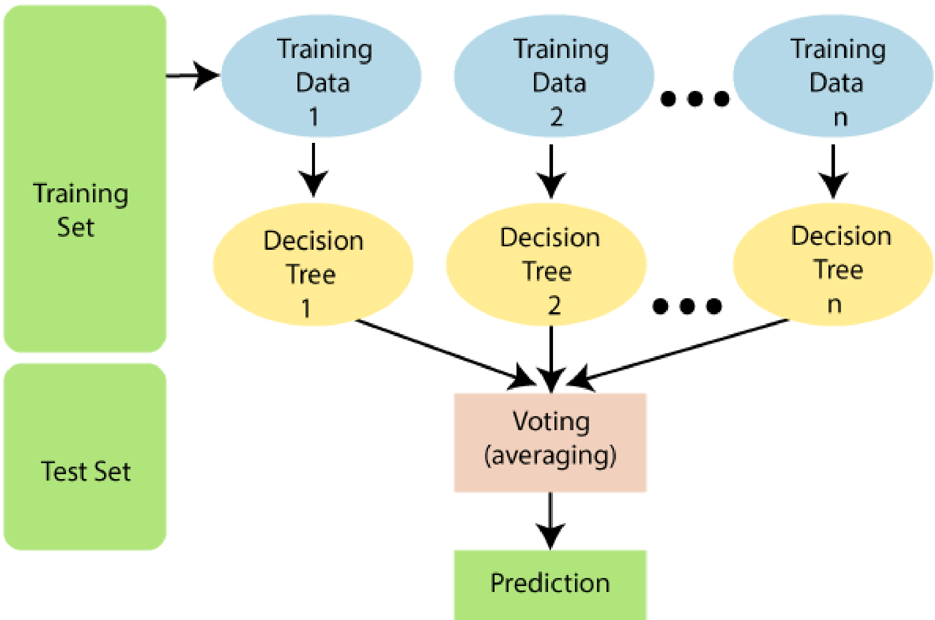

Figure 13. Decision Tree Classifier

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.

Impact of Feature Selection:

Techniques for feature selection were used to find pertinent characteristics for categorization.
The findings imply that feature selection might enhance model performance by minimizing overfitting and concentrating on the most important features, especially when employing techniques of Feature Engineering

Sturdiness and Repeatability:

Data preparation, model training, and assessment protocols were all carefully taken into account during the experimentation phase. Sensitivity analysis demonstrated how consistent the findings were across various dataset setups and parameter settings. Complete documentation of the code and experimental setup was maintained to promote peer review and guarantee repeatability.

Contributions to the Project:

The project advances knowledge of machine learning methods for categorization problems, especially with regard to depression detection and sentimental analysis.
The research provides insights into the advantages and disadvantages of various approaches by assessingmany classifiers and feature selection techniques, assisting practitioners in making well-informed judgments for jobs that are comparable.
To sum up, the study shows how well Decision Tree, Random Forest, Linear SVM, Logistic Regression, Naïve Bayes classifiers work for the job at hand, emphasizes how crucial feature selection is to enhancing model performance, and offers insightful information for potential future research areas in the field.

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.

REFERENCES

[1] G. Shen et al. A multimodal dictionary learning method for depression detection via social media. IJCAI-17 Proceedings: Proceedings of the 26th International Joint Conference on Artificial Intelligence. 2017.

[2] Pedersen T. Utilising lexical decision lists to test Twitter users for sadness and PTSD. From linguistic signal to clinical reality: Proceedings of the Second Workshop on Computational Linguistics and Clinical Psychology2015.

[3] Yazdavar, A.H., Al-Olimat, H.S., Banerjee, T., Thirunarayan, K., & Sheth, A.P. Analyzing clinical depressive symptoms in twitter (2016).

[4] De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. Predicting depression via social media. ICWSM. 13, 1-10 (2013).

[5] Reece, A.G., et al. Forecasting the onset and course of mental illness with Twitter data. Sci. Rep. 7(1),13006 (2017).

[6] Nadeem, M. Identifying depression on Twitter. Preprint at arXiv:1607.07384 (2016).

[7] Shalev-Shwartz, S.; Ben-David, S. Decision Trees. Understanding Machine Learning; Cambridge University Press: Cambridge, UK, 2014.

[8] Kotsiantis, S.B. Supervised machine learning: A review of classification techniques. Informatica 2007, 31, 249–268.

[9] Alloghani, M.A.; Al-Jumeily, D.; Mustafina, J.; Hussain, A.; Aljaaf, A.J. A systematic review on supervised and unsupervised machine learning algorithms for data science. In Supervised and Unsupervised Learning for Data Science; Berry, M., Mohamed, A., Yap, B., Eds.; Springer: Cham, Switzerland, 2020.

[10] Van Engelen, J.E.; Hoos, H.H. A survey on semi-supervised learning. Mach. Learn. 2020, 109, 373–440. [CrossRef]

[11] J. Li and F. Ren 2008. Emotion recognition from blog articles. 2008 International Conference on Natural Language Processing and Knowledge Engineering. pp. 1-8, DOI: 10.1109/NLPKE.2008.4906757.

[12] A. Carrillo-Morales and A. Curiel 2019. Advances Towards the Identification of Mental Disorders Associated with Suicide through Text Processing. 2019 International Conference on Inclusive Technologies and Education (CONTINUE), pp. 121-1217, DOI: 10.1109/CONTIE49246.2019.00031.

[13] Y. Tai and H. Chiu 2007. Artificial Neural Network Analysis on Suicide and Self-Harm History of Taiwanese Soldiers. Second International Conference on Innovative Computing, Information and Control (ICICIC 2007), pp. 363-363, DOI: 10.1109/ICICIC.2007.186

[14] J. Baek and K. Chung 2020. Context Deep Neural Network Model for Predicting Depression Risk Using

Multiple Regression. IEEE Access, vol. 8, pp. 18171-18181, DOI: 10.1109/ACCESS.2020.2968393.

[15] Louis Tay, Psychometric and Validity Issues in Machine Learning Approaches to Personality Assessment: Focus on Social Media Text Mining, European journal of personality.

[16] P. V. Rajaraman Depression Detection of Tweets and A Comparative Test International Journal of Engineering Research & Technology (IJERT) http://www.ijert.org ISSN: 2278-0181 Published by : Vol. 9 Issue 03, March-2020.

[17] Using Content and Activity Features, Hatoon S. ALSAGRI's Machine Learning-Based Approach for 6 Depression Detection in Twitter Information and Systems Transactions, Volume E103-D No. 8, Pages 1825–1832, IEICE.

[18] Muhammad Zulqarnain, Rozaida Ghazali, Yana Mazwin Mohmad Hassim, Muhammad Rehan A comparative review on deep learning models for text classification, Indonesian Journal of Electrical Engineering and Computer Science Vol. 19, No. 1, July 2020, pp. 325~335

[19] Ramin Safa, Peyman BayatAutomatic detection of depression symptoms in Twitter using multimodal analysis Published: 09 September 2021.

[20] A. Zahura and K. A. Mamun 2020. Intelligent System for Predicting Suicidal Behaviour from Social Media and Health Data, 2nd International Conference on Advanced Information and Communication Technology (ICAICT), pp. 319-324, DOI: 10.1109/ICAICT51780.2020.9333463.