Deciphering Depression: Linguistic Analysis of Social Media Data

Manoshika Catherine S J, *Department of CSE, Karunya Institute of Technology and Sciences, Coimbatore, India*
@manoshikacatherine@gmail.com

***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 common words used by a depressive person and the words used 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.**

***Index Terms*—Machine Learning, Depression Detection, Stress, Emotions, Twitter, Emotions, Natural Language Processing**

I. Introduction

Due to the fluctuating nature of mental states, it can be difficult to distinguish non-depressed users on online social networking platforms. People may have a wide range of emotions when they publish on the internet, communicate with others, and engage in other online activities. Since social media is so widely used these days, especially on sites like Twitter, which is popular in many developing countries, involvement in social media networks has skyrocketed. Twitter is a place where people can openly debate a wide range of issues and share their thoughts, opinions, and accomplishments.

The ability to express oneself freely on the internet, however, can also result in increased emotional reactions and mental health issues, which is similar to addiction-like behavior. People who communicate with others online may suffer negative consequences for their mental health, so it's important to keep an eye out for symptoms of psychological discomfort in shared messages and photographs. It's critical to identify and treat possible mental health problems early on before they get worse.

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.

To potentially identify distinctive characteristics and patterns, deep learning algorithms are utilized. The co-occurrence of anxiety and depression can have severe consequences, as they share symptoms and traits common to depressive disorders. In our analysis, we examine both bigrams and unigrams in Twitter data to capture feature spaces. For example, the tweet "I Love Kindle, It's Amazing" yields unigrams such as "I, Love, Kindle, Its, Amazing" and bigrams like "I Love, Love Kindle, Kindle Its, It's Amazing." Additionally, we consider the term frequency representation (count of term occurrences in a document) and term presence representation (presence or absence of a term in a document) in the dataset.

The testing dataset is utilized for both training and testing purposes, focusing on identifying individuals who frequently post depressive content or engage in related online behaviors. Moreover, the relationship between depression intensity and cardiovascular disease risk can be assessed based on social media activity.

Twitter's widespread usage provides users with a platform to express their opinions while maintaining online connections. The discussion addresses three separate disorders: anxiety

[7] two tests are utilized, and emotional connection is investigated using supervised machine learning classifiers. They used categorization techniques to categorize depression.- pertinent posts on social media.

The study recommended in [1] combines emotional feature analysis with N-gram language modeling to assess anxiety levels and build a classifier for identifying clinical depression based on behavioral features extracted from Twitter data. Another study explored emotions using Facebook data and developed a statistical model for emotion forecasting

This suggests that depression can be predicted using linguistic characteristics, utilizing the Twitter dataset for training a decision list to detect depression. They also propose a method for developing social media classifiers for post-traumatic stress disorder, demonstrated using Twitter data..

In [1], authors review research on using online social networks for public health forecasting, using Twitter data to make predictions based on users' tweets, status updates, social connections, the timing of usage, and overall behavioral patterns..

Researchers in [5] explored various statistical techniques and machine learning algorithms to accurately identify depression in individuals, analyzing prediction rate variations using the Twitter dataset. They emphasize the importance of each tweet's persuasive ability in sensitive situations.

[6] conducted an analysis of Twitter data, focusing on identifying emotions related to psychological disorders or mental health issues. Employing sentiment analysis techniques, they categorize emotional states based on subjects' behavioral characteristics..

disorder, characterized by susceptibility to emotional distress from conflicting online interactions; anxious depression, observed in individuals who actively share their emotions online; and social media's role in facilitating psychological support for anxiety and depression.

II. METHODOLOGY

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

A. 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).

B. 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. The One-Hot approach is used to parse the stemmed input text to extract features from these input words. In a machine learning prediction model, the traits are binary patterns that can be utilized to represent depression.

C.*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. Using a variety of sources, including text (sentiment analysis of social media posts or clinical notes), physiological signals (heart rate variability, sleep patterns), behavioral data (activity levels, social interactions), and demographic data (age, gender), 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.

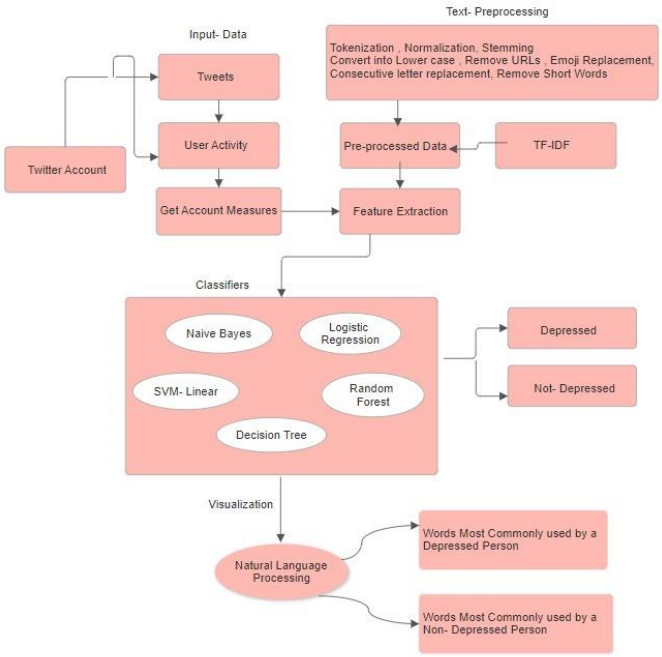


Figure 1. *Methodology for Text-based Depression Detection*

D. *Machine Learning Models*

1) **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

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

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

2) **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.

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

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.

3) **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.

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

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. 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.

4) **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.

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, rgression, and classification

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

5) **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 [10] 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.

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

6) **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.

III. RESULT AND CONCLUSION

In this work, the detection of depression from social media, tweets from Twitter [12]. Comparing the accuracy of the dataset validation to other existing individual classifiers, it is more accurate.

The Feature Engineering for choosing, modifying, and converting unprocessed input into features for supervised learning by the Natural Language Processing (NLP) Techniques.

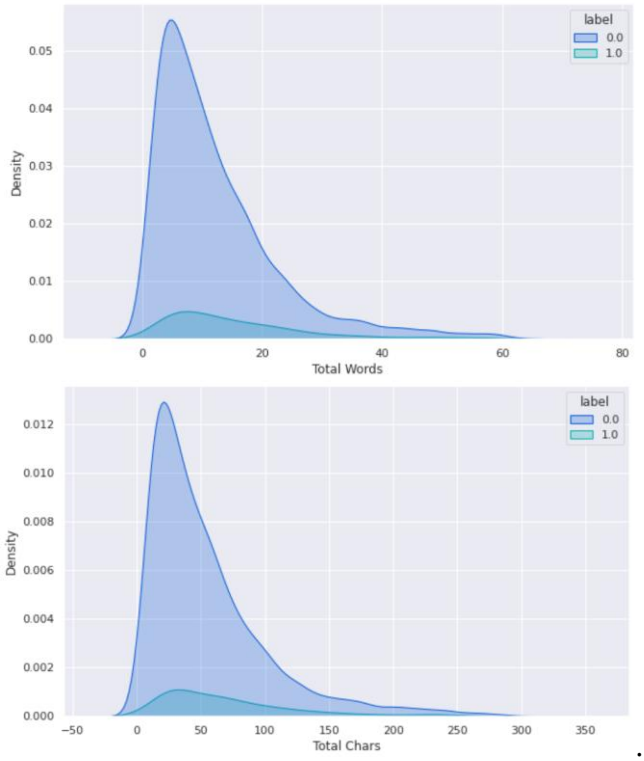


Figure2. *Feature Engineering*

The Processing of the Text is done by converting all the text into lowercase letters, removing the URLs, removing the Punctuations, and removing the Stop Words. The stemming, Tokenization, and Transformation of the data are done successfully.

The words that are used frequently by the Depressed Person and the Non-Depressed Person are determined and visualized using the Natural Language Processing(NLP) Techniques using the Viridis Colormap.

Figure 3. *Visualization of Non- Depressive Words*

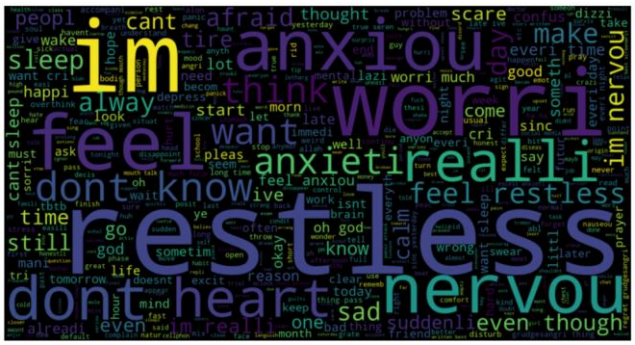


Figure 4. *Visualization of Depressive Words*

The Most Commonly used Words used by a Person when Not Depressed and when Depressed has been predicted and Visualized using Matplotlib and Seaborn

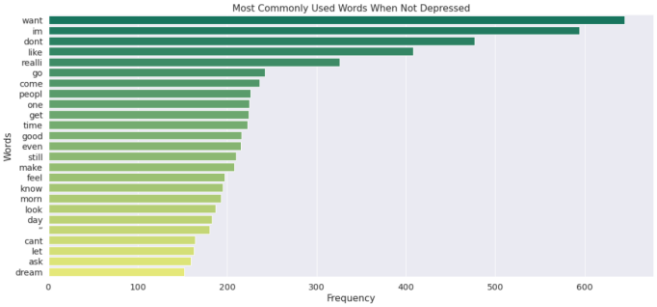


Figure 5. *Most Commonly used Words When Not- Depressed*

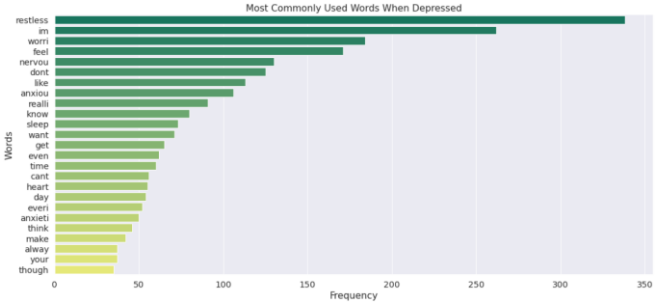


Figure 6. *Most Commonly used Words When Depressed*

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.

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

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.

Utilizing the confusion matrix (CM), standard accuracy (Acc), precision (P), recall (R), and F1 scores, the viability of automated prediction will be examined and cross-validated [13–15].
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.

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

Precision is defined as the proportion of correctly classified positives to all predicted positives. It aims to ascertain the percentage of users who are labeled as depressed who, under our circumstances, are indeed depressed.

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

Recall is defined as the proportion of accurately identified positives to all positives. The objective in this instance is to determine the percentage of all users who are diagnosed with depression accurately.

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

The F1 Score, sometimes referred to as the F-measure, is widely used as a classification evaluation metric since it equally weights each metric as the harmonic mean of recall and precision.

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

Confusion Matix: Naïve Bayes

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

Confusion Matix: Random Forest

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

Confusion Matix: Linear SVM

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

Confusion Matix: Logistic Regression

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

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

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 the algorithms that are implemented in this paper.

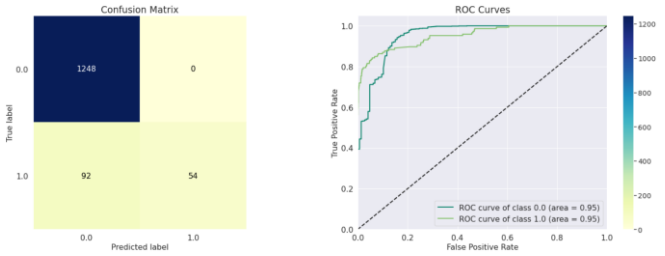


Figure 7. *Confusion Matrix and Roc Curve for Naïve Bayes*

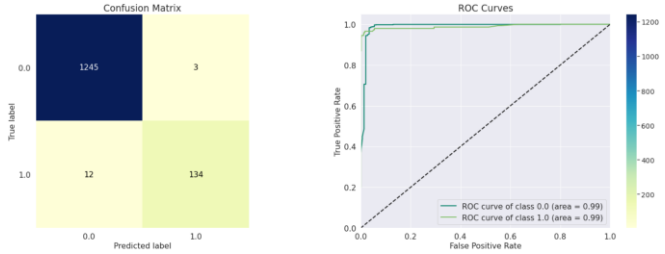


Figure 8. *Confusion Matrix and Roc Curve for Random Forest*

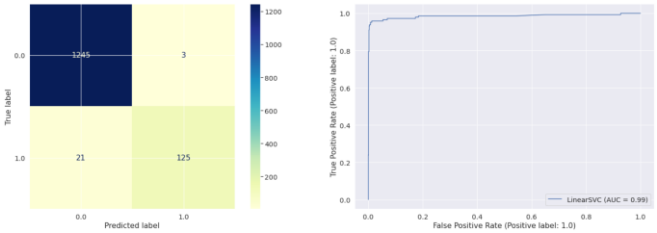


Figure 9. *Confusion Matrix and Roc Curve for Linear SVM*

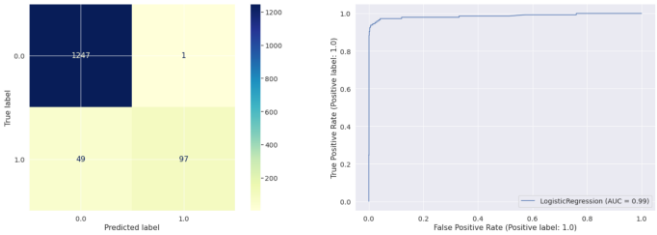
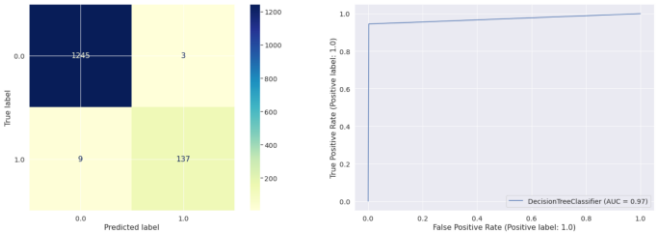


Figure 10. *Confusion Matrix and Roc Curve for Logistic Regression*



. ***Figure 11.*** *Confusion Matrix and Roc Curve for Decision Tree*

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 accurately 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.

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.

IV. *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: A 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.