Analyzing the Shadows: Machine Learning Approaches for Depression Detection on Twitter

Fiza Azam[1], Memoona Sami[2], Maha Agro[3], Amirita Dewani[2]
[1] National University of Sciences & Technology (NUST)
[2] Mehran University of Engineering & Technology Jamshoro
[3] Muhammad bin Zayed University of Artificial Intelligence UAE
*Correspondence. Memoona Sami, Email. sw.memoona@gmail.com

Citation | Azam. F, Agro. M, Sami. M, Dewani. A, "Analyzing the Shadows. Machine Learning Approaches for Depression Detection on Twitter", IJIST, Vol. 6 Issue. 3 pp 1402-1416, Sep 2024
DOI | https://doi.org/10.33411/ijist/20246314021416
Received | Aug 5, 2024 **Revised |** Sep 10, 2024 **Accepted |** Sep 14, 2024 **Published |** Sep 16, 2024.

Depression is a leading cause of disability worldwide, affecting approximately 4.4% of the global population. It can escalate from mild symptoms to severe outcomes, including suicide, if not treated early. Thus, developing systematic techniques for automatic detection is crucial. Social media platforms like Facebook, Twitter, TikTok, Snapchat, and Instagram provide users with the means to share personal feelings and daily activities, offering valuable insights into their thoughts and behaviors. This research aims to identify users who publicly disclosed their diagnoses and collect their data from Twitter. We created three different datasets, each varying in the number of tweets stored based on criteria discussed later. We selected six classifiers for analysis. Support Vector Machine (SVM), Logistic Regression, Random Forest, Max Vote Ensemble, Bagging, and Boosting. We conducted two analyses. In the first, textual data was converted into embeddings using the Bag of Words approach before analysis. In the second, a multivariate analysis, we trained algorithms on multi-dimensional data. Our findings revealed that Logistic Regression outperformed other techniques on smaller datasets. However, the Boosting algorithm yielded the best results on a dataset of 3,200 tweets, and the Bagging algorithm excelled when trained on 3,200 tweets of multivariate data. Overall, nearly all algorithms performed well on the 3,200-tweet datasets.

Keywords. Depression, Machine Learning, Major Depressive Disorder, Sentiment Analysis, SVM, Random Forest.

Introduction.

Major Depressive Disorder (MDD), commonly known as depression, is one of the most severe psychiatric disorders worldwide [1], making it the leading cause of disability globally [2]. A study disclosed that approximately 322 million people suffer from MDD, contributing to 4.4% of the global disease burden [3]. According to the World Health Organization, depression is more common in developing countries, where an estimated 10-44% of the population has experienced it, and approximately 50.8 million people are currently affected. In 2021, an estimated 280 million people, including 5% of all adults, experienced depression [4]. Depression is often associated with poor socioeconomic status, with individuals from disadvantaged backgrounds being more likely to suffer from this disorder [5]. A study [6] identified factors such as academic stress, teenage marriages, limited job opportunities, unemployment, drug abuse, restricted access to higher education, and political violence as contributing factors to the high prevalence of depression among young adults. Additionally, age plays a significant role, with younger individuals being more vulnerable to depression [7], while the Centers for Disease Control and Prevention, report a 1% to 5% prevalence among senior citizens [8]. The occurrence of depression is also influenced by gender, with women being more prone than men, as 1 in 8 women may experience MDD at some point in their lives [9].

Early detection of depression is crucial for improving quality of life, reducing the risk of suicide, and decreasing drug use. The Diagnostic and Statistical Manual of Mental Disorders (DSM-V) defines a major depressive episode as a persistently dejected mood and lack of interest in activities for at least two weeks, along with six other symptoms, including sleep disturbances, changes in weight and appetite, fatigue, concentration problems, psychomotor agitation/retardation, and recurrent suicidal thoughts [10]. Diagnosing depression is challenging due to its overlap with other illnesses and varying symptoms [11]. Physicians often overlook depression unless major symptoms are reported, and the link between depression and somatic complaints is frequently missed [12]. Traditionally, depression levels are measured using questionnaires such as CES-D [13], BDI [14], PHQ-9 [15], and SDS [16]. However, these self-reported or third-party-filled forms can be unreliable and subject to manipulation, highlighting the need for a systematic diagnosis mechanism.

Recently, social media platforms such as Facebook, Twitter, Reddit, and Tumblr have provided a way to observe behavioral traits linked to thoughts, social interactions, moods, and activities. Women tend to be especially active and expressive on social media platforms, where users frequently share their emotions and daily experiences. The language used can reflect feelings of guilt, shame, and self-deprecation, characteristic of major depression [17]. Thus, social media can serve as a valuable tool for early depression diagnosis. Using natural language processing techniques, signs of depression can be detected from online data. The primary goals of this study are to identify depressed populations on social media, collect their data, and use it to train machine learning models such as support vector machines, logistic regression, random forests, and ensemble methods like bagging and boosting for accurate MDD identification.

The paper is structured as follows. The Introduction provides the contextual background and defines the research problem. The Literature Review discusses existing studies and identifies research gaps. The Objectives section outlines the specific goals of this research. The Research Methodology details the study's phases and processes. The Classification Metrics section describes the experimental setup and evaluation criteria. The Results section presents a comparison of outcomes across different tweet datasets. The Discussion focuses on model interpretability and their potential for real-time application. Finally, the Conclusion summarizes the key findings and offers suggestions for future research directions.

Literature Review.

The methodology proposed by [18] involved collecting data by scraping tweets from Twitter. The query “I was diagnosed with X disease” was used, with "X" being replaced by

"depression" and "PTSD." This data was then examined using various machine-learning techniques to detect the presence of these disorders. Similarly, a previous study [17] used Twitter's public API to collect data with the query "I was diagnosed with X." The experiments concluded that language models produced better results than LIWC.

Research by [19] formulated a dataset of 140,946 tweets from about 90 people, identifying 1,000 tweets with depressive language. They noted the high-frequency usage of first-person pronouns and negative language among people suffering from depressive disorders. S. Tsugawa et al. collected SDS levels from 50 Japanese candidates and then acquired tweets from each participant to assess their depression levels. Using a regression model, they found a positive association between Zung's depression scale and the results obtained [20]. Another study [21] fetched data from Twitter using the keyword "Depression" and analyzed it through multiple regression and LIWC models. The findings revealed that individuals with depression frequently share their diagnosis online; however, gender did not emerge as a significant factor. Researchers in [2] collected Spanish and Portuguese tweets indicating depression, pregnancy, flu, or eating disorders. The author [22] found 3000 tweets of users showing signs of depression or PTSD by querying "I was just diagnosed with depression or PTSD." While no significant results were found, the research provided insights for future work. The author [23] used CNN, RNN, SVM, and other deep learning paradigms to identify depression on Twitter, with the CNN model performing the best. Kabir et al. [24] introduced a typology for classifying social media texts to detect varying levels of depression severity. The results demonstrate that the proposed framework effectively enhances the accuracy of depression detection, offering a valuable tool for mental health assessment.

F. Cacheda et al. gathered Reddit data and processed it to obtain tuples containing ID and writing, employing features like textual similarity and semantic similarity for early depression diagnosis. The author [25] used crowdsourcing methods to gather Twitter data from users scoring significantly on the CES-D, findings revealed that depressed people often have tightly-knit egocentric social graphs and use negative language. Another study [26] analyzed tweets to detect signs of depression using six machine learning algorithms, concluding that SVM outperformed the others. A different study compared baseline LSTM and a hybrid model combining BiLSTM and CNN, revealing that the hybrid model generated better results [27].

This paper [28], proposed a methodology for creating a dataset from social media data, classifying it into not depressed, moderately depressed, and severely depressed categories. Data augmentation techniques were used to address data imbalance, and several traditional machine learning algorithms were applied. The study found that a model trained with Word2Vec embeddings and a Random Forest classifier achieved the best results with an accuracy of 0.877 and a high F1 score [28]. Research in [29] attempted to label Reddit comments as "depressed" or "non-depressed" using traditional machine learning and deep learning techniques, with SVM, Logistic Regression, CNN, and BERT models. Another study [30] deployed several traditional machine learning models to distinguish between suicidal and non-suicidal posts on Reddit, with SVM generating the best results. Akinyemi et al. [31] showed that machine learning techniques, especially ensemble methods, are effective in detecting and classifying cyberbullying in social media texts. The author [32] gathered Facebook data to assess users' depression levels, employing SVM, K-Nearest Neighbors, Decision Trees, and Ensemble methods. Research [33] analyzed Facebook status updates of patients from certain healthcare facilities to detect depression. A study [34] used RNN and LSTM techniques to detect depression using a Kaggle dataset, achieving 99% accuracy and outperforming traditional machine learning models. Another study [35] analyzed real-time tweets using sentiment analysis techniques and a neural network, with a Bi-directional LSTM model achieving 90% accuracy. The study [36] proposed the automatic detection of depression using an explainable Multi-Aspect Depression Detection with a hierarchical attention Network (MDHAN), outperforming baseline models.

Research [37] used natural language processing and sentiment analysis to detect depression on Twitter, finding the Random Forest classifier outperformed SVM. A paper [38] reviewed studies identifying depressive mood disorder on social networks using sentiment and emotion analysis techniques. Another review [31] examined the papers diagnosing depressive disorders through social media, suggesting advances in NLP are needed for finer granularity and addressing the ethical implications of privacy breaches. The paper examined social media's impact on adolescents' mental health, reviewing literature across four domains. time spent, activity, investment, and addiction. They found correlations between social media and mental health but no significant causation. A review identified relationships between social networking sites and their effects on depression and anxiety, noting mixed results and the need for longitudinal studies to better understand correlations and causation.

Objectives.

- To Develop a robust methodology for accurately identifying Twitter users who exhibit signs of depression, distinguishing them from a control group of non-depressed users.
- Implement a comprehensive data extraction process to collect relevant behavioral and linguistic data from the profiles of the identified user groups.
- To Create three distinct datasets encompassing different temporal spans (two weeks, one month) and tweet volumes (up to 3,200 tweets), incorporating multivariate data to facilitate thorough analysis.
- Train and fine-tune machine learning classifiers on the constructed datasets to predict depression in Twitter users, based on their word usage patterns and age.
- To Conduct thorough testing and validation of the trained models to assess their predictive accuracy and overall performance, ensuring their effectiveness in real-world applications.

Methodology.

Data Collection.

This section outlines the methodology for identifying depressed and control candidates on the Twitter platform using an API. Initially, potential candidates were selected based on specific criteria, and their data was then scraped to create datasets. The data was organized into three distinct datasets, each varying in size. Once the datasets were prepared, they were subjected to a data cleaning process as illustrated in Figure. 1.

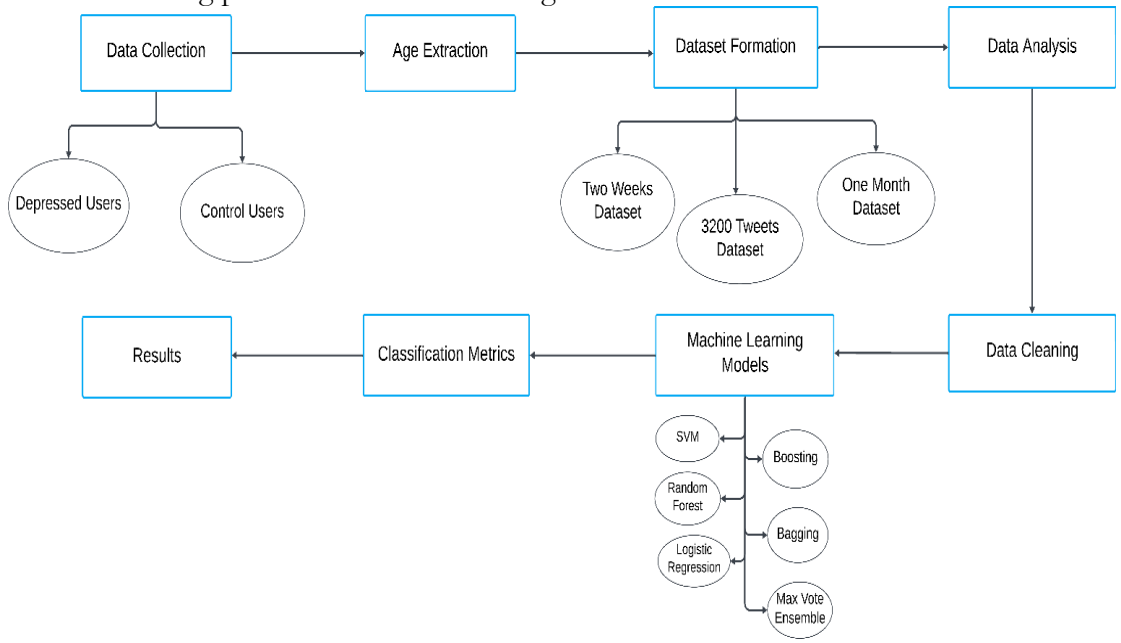


Figure 1. Diagram of the Complete Proposed Research Methodology

Detection and Analysis of Depressed Users.

The identification of users exhibiting signs of depression was achieved using a methodology proposed by [17], bypassing traditional mental health questionnaires. Social media platforms, such as Twitter, serve as venues where individuals often share their mental health conditions openly. Users frequently disclose their mental health diagnoses through statements such as "I was diagnosed with X," where X represents various mental health conditions [22]. To gather data on users with depression, two specific queries were on Twitter.

- "I was diagnosed with depression"
- "I am diagnosed with depression"

For each query, 2,000 tweets were collected, resulting in a total of 4,000 tweets. These tweets underwent manual verification to ensure authenticity. Table 1 presents examples of both genuine and non-genuine tweets identified through this process. The manual verification focused on confirming the authenticity of tweets indicating a depression diagnosis or history. This process resulted in 400 verified tweets, which were further screened under the following conditions.

- The user must have posted at least 100 tweets.
- A significant proportion of the tweets should be in English.

Out of the initial 400 tweets, 378 met the screening criteria and were included in the subsequent analysis phase.

Control Users.

A control dataset was obtained using the query "Today is my birthday" to generate a set of random tweets. A total of 3,200 tweets were collected using this query. The selection process for the control group followed the same criteria applied to the depressed user data. each candidate had to have posted at least 100 tweets, with a high proportion in English. Following this, each tweet was filtered for the presence of terms related to depression using a pipeline similar to [19], incorporating 15 relevant keywords. Depression, Anxiety, Distressed, Demotivated, Insomnia, Lonely, Empty, Exhausted, Worried, Overwhelmed, Tired, Sad, Discouraged, Cry, and Nervous. Candidates whose tweets contained any of these terms were excluded from the control set. Consequently, 3,080 tweets from the control group were retained for further analysis.

Age Extraction.

Age is a significant factor in depression studies [20][21]. User profiles that progressed to the next phase were manually examined for age information. Due to the limited disclosure of age on profiles, the Lexicon tool [34] was employed for age prediction. For users without available age details, their most recent 100 tweets were analyzed using the tool to estimate their age.

Dataset Formation.

Three datasets were constructed, each varying in tweet count based on temporal constraints. The tweet counts for each dataset are illustrated in Figure. 2.

- **Two-Week Dataset.**

Tweets from depressed users were collected from the date of their depression diagnosis up to two weeks later. Similarly, for control users, tweets were gathered from the day of their birthday tweet up to two weeks thereafter. The resulting datasets contained 82,077 tweets for depressed users and 7,699 tweets for control users.

- **One-Month Dataset.**

For both groups, tweets were collected from the date of the diagnosis or birthday tweet up to one month later. The dataset comprised 15,6165 tweets from depressed users and 12,815 tweets from control users.

- **3,200 Tweets Dataset.**

Due to API limitations, the dataset was restricted to a maximum of 3,200 tweets per user. The final datasets included 404,643 tweets from depressed users and 103,325 tweets from control users.

Anonymization.

To protect user privacy, all usernames were replaced with numeric identifiers, and any personal information such as geolocation data and URLs was removed from the datasets.

Data Structure.

Six SQL tables were created to organize the data, including depress_week_data, control_week_data, depress_month_data, control_month_data, depress_3200_data, and control_3200_data. Each table included the following columns.

- **Id.** Unique identifier for each user.
- **Tweets.** Concatenated tweets from a user.
- **Age.** Age of the user.
- **Count.** Total number of tweets.
- **Emoji Count.** Total number of emojis and emoticons.
- **Noun Count.** Total number of nouns used.
- **Verb Count.** Total number of verbs used.
- **Pronoun Count.** Total number of pronouns used.
- **Personal Pronoun Count.** Total number of personal pronouns used.

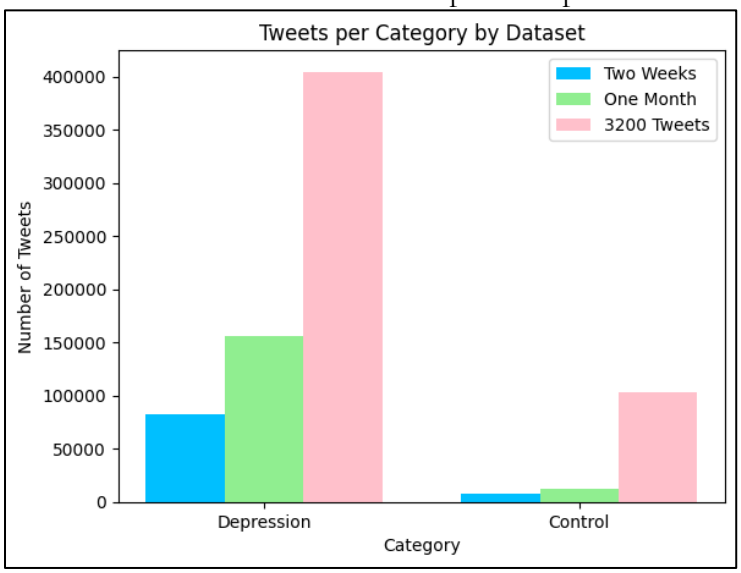


Figure 2. Tweet Count Per Each Dataset

Data Analysis.

Analysis of the 3,200-tweet dataset revealed that the average age of depressed users was approximately 25 years, compared to 29 years for control users. Depressed users averaged 123 emojis per tweet, whereas control users averaged 32. Depressed users used an average of 5,551 nouns and 5,188 verbs, while control users used 867 nouns and 686 verbs. Pronoun and personal pronoun usage averaged 2,593 and 1,969, respectively, for depressed users, compared to 511 and 189 for control users.

Data Cleaning.

The data cleaning process involved.

- Removal of URLs using regular expressions via regex version 2022.3.2.
- Replacement of emojis and emoticons with textual equivalents using emote API version 3.1.

- Deletion of non-English words using NLTK version 3.8.
- Removal of stop words except first-person pronouns, as they were deemed significant in identifying depressed users [19].
- Tokenization and lemmatization to standardize the text data.

Following cleaning, a bar chart was created to visualize the frequency of words in the depressed dataset, highlighting the prevalence of first-person singular pronouns and emoticons shown in Figure. 3.

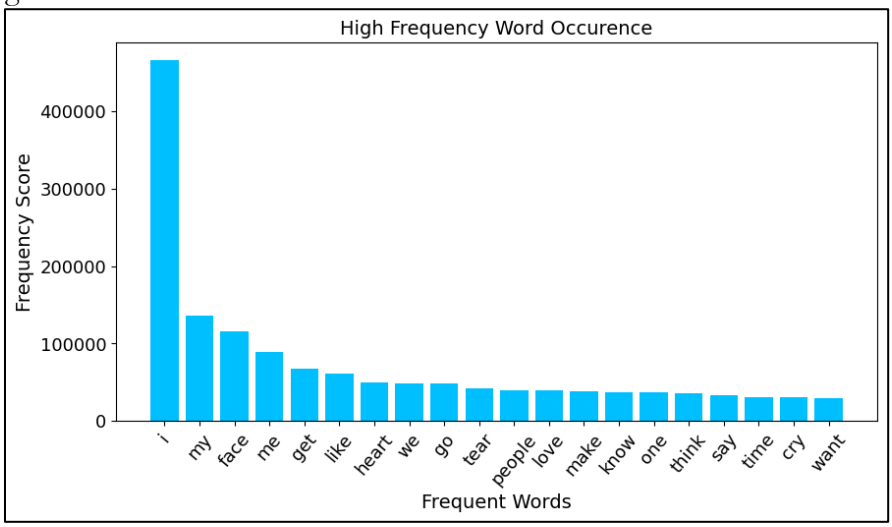


Figure 3. Word Frequency in the Depression Dataset

Data Limitations and Biases.

- The study's data represents only English-language tweets that were analyzed, thus lacking in representing the broader Twitter user population.
- The detection method does not confirm a formal diagnosis of depression but identifies users who self-report such conditions.
- Some control users may be undiagnosed or unwilling to disclose their mental health status.
- Twitter users may not be representative of the global population.
- A significant number of tweets from the control group were lost during the data cleaning process, leading to an imbalance between the depression and control datasets. This imbalance could introduce potential bias during model training.
- We visualized age distribution for all selected candidates (both depressed and control) by categorizing it into ranges, as shown in Figure 4. This revealed a noticeable gap in data for individuals aged 35 to 67, likely due to lower Twitter usage among people in this age group.

Machine Learning Algorithms.

- **Support Vector Machine (SVM).** This approach was utilized to create a decision boundary in n-dimensional space to separate depressed and control data points. A linear SVM was used, aiming to maximize the margin between classes. To enhance model performance, a grid search was conducted to optimize parameters, including gamma set to auto, cache size to 12,000, and max_iter to -1.
- **Logistic Regression (LR).** Employed for binary classification, estimating event probabilities, and using a sigmoid function for binary outcomes. A grid search was conducted, resulting in the selection of the optimal regularization parameter (C) set at 1 and the L2 regularization type for improved model performance.

- **Random Forest (RF).** An ensemble method combining multiple decision trees to improve classification accuracy. Through grid search, the model was optimized to include 150 decision trees, using the Gini criterion to assess splitting quality.
- **Max Vote Ensemble (MVE).** Aggregated predictions from multiple classifiers, including Logistic Regression, SVM, Random Forest, Extreme Gradient Boosting (XGB), and Decision Tree, using hard voting to determine the final classification. Each classifier's parameters were fine-tuned using grid search to optimize performance.
- **Bagging (BAG).** An ensemble technique that reduces dataset variance by training multiple weak classifiers on different subsets of data. A Decision Tree Classifier was used as the base classifier, with 500 trees. All other parameters were optimized through grid search.
- **Boosting (BST).** An ensemble method that sequentially trains classifiers to correct errors made by previous models. The base classifier was Gradient Boosting, and its hyperparameters were selected through grid search to optimize performance.
- **Bag Of Words (BOW).** Applied to convert textual data into a vector representation, capturing unigram word frequencies. This method transforms each document into a vector of word occurrences.

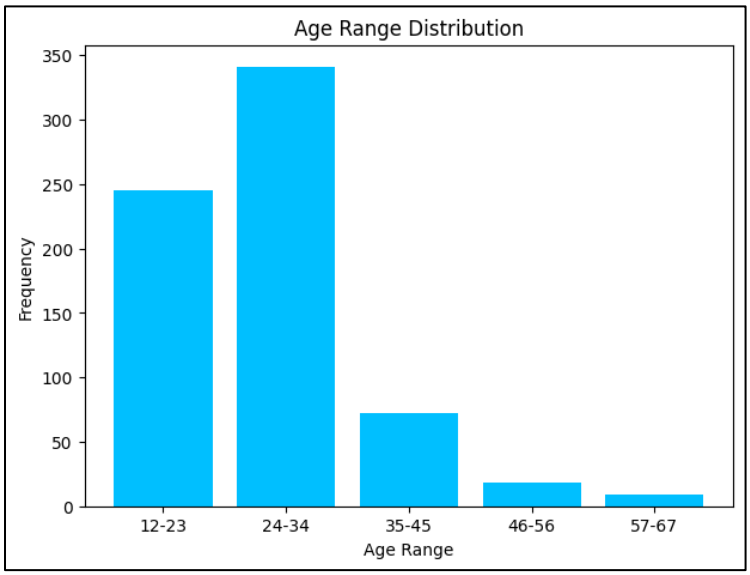


Figure 4. Age Range Distribution of Twitter Users.

Classification Metrics.

The efficacy of each algorithm was evaluated using three distinct datasets and analyzed across five classification metrics. These metrics were derived from the counts of True Positives, True Negatives, False Positives, and False Negatives. For this study, these terms are defined as follows.

- **True Positive (TP)**. Instances where the depressed class was correctly identified as depressed.
- **True Negative (TN)**. Instances where the control class was correctly identified as control.
- **False Positive (FP)**. Instances where the control class was incorrectly classified as depressed.
- **False Negative (FN)**. Instances where the depressed class was incorrectly classified as control.

The classification metrics used in this analysis are.

- **Precision**. Precision quantifies the accuracy of the positive predictions made by the classifier. It is defined as the ratio of True Positives to the sum of True Positives and False Positives.

\text {Precision} = \frac {TP} {TP + FP}

This metric reflects how many of the instances labeled as depressed by the classifier are truly depressed.

- **Recall.** Recall measures the proportion of actual positive instances (depressed users) that were correctly identified by the classifier. It is calculated as.

\text {Recall} = \frac {TP} {TP + FN}

This metric indicates the effectiveness of the classifier in identifying all relevant instances of the depressed class.

- **F1-Score.** The F1-Score is the harmonic mean of Precision and Recall, providing a single metric that balances the trade-off between them. It is defined as.

\Text{F1-Score} = 2 \times \frac {\text {Precision} \times \text {Recall}} {\text {Precision} + \text {Recall}}

This score is useful for evaluating models where both Precision and Recall are important.

- **Accuracy.** Accuracy measures the proportion of correctly classified instances (both depressed and control) among the total number of instances. It is computed as.

\text {Accuracy} = \frac {TP + TN} {TP + TN + FP + FN}

This metric provides an overall measure of the classifier's performance.

- **Area Under the Curve (AUC)**. The AUC of the Receiver Operating Characteristic (ROC) curve represents the classifier's ability to distinguish between the two classes (depressed and control). It provides an aggregate measure of performance across all classification thresholds. An AUC value closer to 1 indicates better separability between the classes.

Results.

The datasets, encompassing two-week, one-month, and 3200-tweet intervals, were consolidated into three comprehensive datasets. Each dataset integrated data from both depressed and control categories. The analysis was conducted in two primary phases. The first phase involved a textual analysis where tweets were utilized as features, with the depression status (1 for depressed and 0 for control) serving as the label. In this phase, the Bag-of-Words (BOW) approach was employed to convert the tweets into vector representations, enabling the application of machine learning algorithms to the textual data.

In the second phase, a multivariate analysis was conducted on a dataset comprising 3,200 tweets, incorporating multiple features alongside textual data. These features included age, total tweet count per user, and counts of emojis, nouns, verbs, pronouns, and personal pronouns. Before algorithm training, the data underwent min-max normalization to ensure consistent scaling across features. This normalization process adjusted the feature values to a uniform range, facilitating the convergence of the algorithms and enhancing their performance. These analyses aimed to assess the efficacy of various features and data representations in accurately classifying tweets based on depression status. The results are detailed in the next sections. Table 1 presents the performance metrics for each classifier trained on the dataset, while Figure 5 visualizes the Area Under the Curve (AUC) for each model, highlighting their comparative effectiveness in predicting depression status.

Two Weeks of Data.

The experimental results for classifying user sentiments revealed the distinct performance characteristics across various algorithms. Support Vector Machine (SVM) exhibited high precision but the lowest recall among the classifiers, achieving an accuracy of 0.73 and an Area Under the Curve (AUC) score of 0.74, which were the lowest in comparison to

other models. Logistic Regression outperformed other algorithms, demonstrating precision and recall scores of 0.88 and 0.85, respectively. Additionally, it achieved accuracy, F1 score, and AUC scores of 0.85, 0.86, and 0.85, respectively. Random Forest also performed well, with precision and F1 scores of 0.83 and 0.82, respectively. The Max Vote Ensemble method delivered the second-best results, with precision and recall values of 0.86 and 0.83, and accuracy and AUC scores of 0.83. The Bagging algorithm achieved precision and recall scores of 0.85, and an accuracy of 0.84, with its AUC score matching that of the Max Vote Ensemble. Finally, the Boosting algorithm recorded precision, recall, and F1 scores of 0.81, but its accuracy and AUC were 0.80 and 0.79, respectively.

One Month of Data.

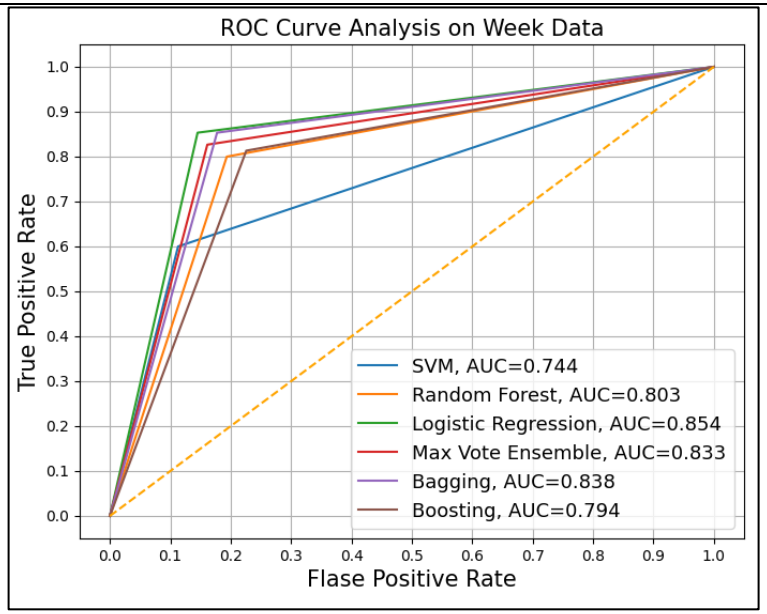
The experimental results highlight significant variations in performance metrics across different algorithms. The Support Vector Machine (SVM) yielded the lowest recall and F1 scores of 0.57 and 0.68, respectively, while achieving precision, accuracy, and AUC scores of 0.86, 0.72, and 0.72. In contrast, Logistic Regression emerged as the best-performing model with an AUC, F1 score, and accuracy of 0.80, along with precision and recall scores of 0.86 and 0.74, respectively. The Random Forest classifier produced precision and recall scores of 0.80 and 0.70, with AUC and accuracy scores of 0.74, and an F1 score of 0.75. The Max Vote Ensemble achieved precision and recall scores of 0.83 and 0.74, with accuracy, F1, and AUC scores of 0.78, 0.79, and 0.78, respectively. The Bagging ensemble attained precision, recall, and F1 scores of 0.83, 0.77, and 0.80, with accuracy and AUC scores of 0.79. Finally, the Boosting algorithm delivered precision, recall, accuracy, F1, and AUC scores of 0.79, 0.81, 0.78, 0.80, and 0.77, respectively.

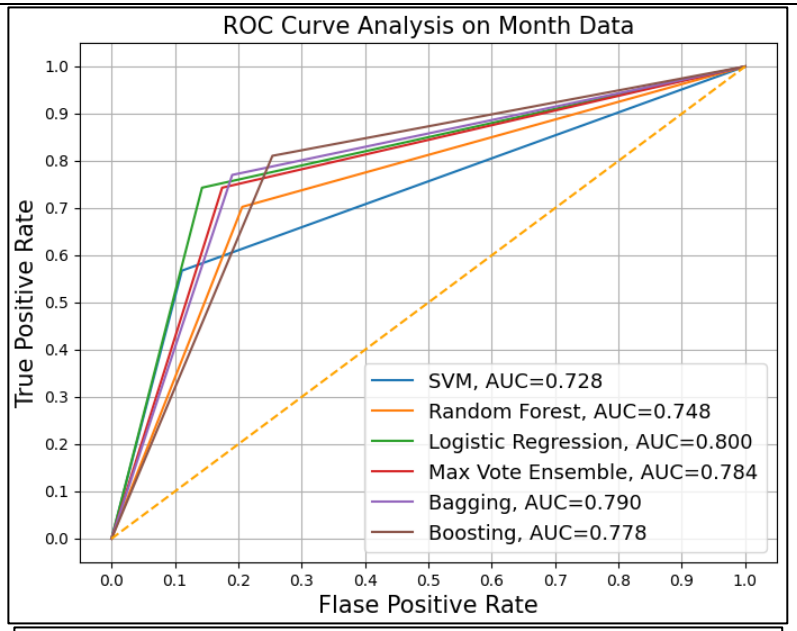
3200 Tweets Data.

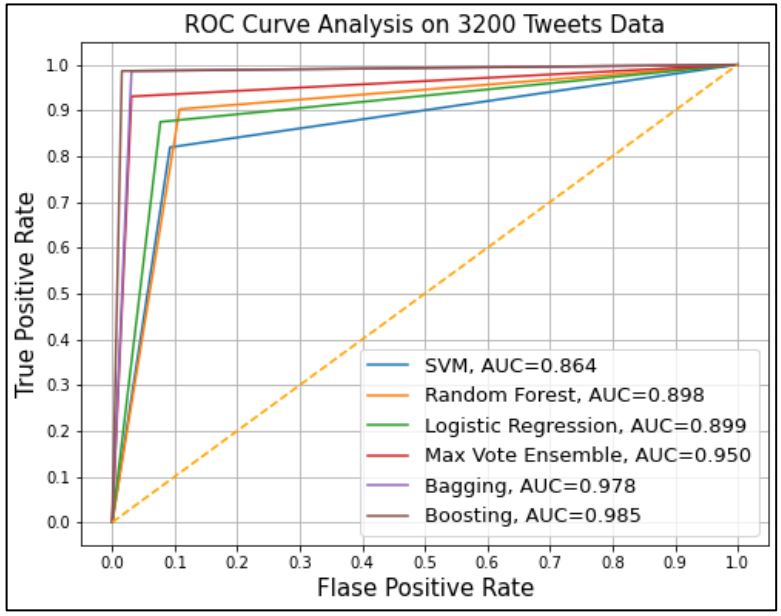
The outcomes obtained using the Support Vector Machine (SVM) on this dataset showed significant improvements compared to other datasets. The algorithm achieved precision and recall scores of 0.91 and 0.82, respectively, with accuracy, F1, and AUC scores all reaching 0.86. Logistic Regression also performed well, generating precision, recall, and AUC scores of 0.93, 0.88, and 0.89, respectively, with accuracy and F1 scores both at 0.86. The Random Forest classifier produced precision, recall, accuracy, and F1 scores of 0.90, while its AUC score was 0.89. The Max Vote Ensemble achieved high performance with precision and recall scores of 0.97 and 0.93, and accuracy, F1, and AUC scores of 0.95. The Bagging ensemble excelled with precision and AUC scores of 0.97, accuracy and F1 scores of 0.98, and a recall score of 0.99. Notably, the Boosting algorithm outperformed all others on this dataset, with precision, recall, accuracy, and F1 scores of 0.99, and an AUC score of 0.98.

3200 Tweets Multivariate.

The analysis conducted on the largest dataset, comprising 3200 tweets, yielded the following results. The Support Vector Machine (SVM) achieved precision, recall, accuracy, F1, and AUC scores of 0.92, 0.68, 0.79, 0.78, and 0.79, respectively. Logistic Regression produced precision and recall scores of 0.92 and 0.73, with accuracy, F1, and AUC scores around 0.82, 0.81, and 0.84. The Random Forest classifier achieved a precision score of 0.87, recall of 0.84, accuracy of approximately 0.85, F1 score of 0.86, and an AUC of about 0.82. The Max Vote Ensemble yielded precision and recall scores of 0.89 and 0.78, with all other metrics scoring around 0.83. The Bagging Ensemble outperformed other methods with a precision score of approximately 0.90, recall of 0.85, accuracy and AUC scores of around 0.87, and an F1 score of about 0.88. Lastly, the Boosting algorithm produced precision, recall, and F1 scores of 0.86, with accuracy and AUC scores both at 0.85.







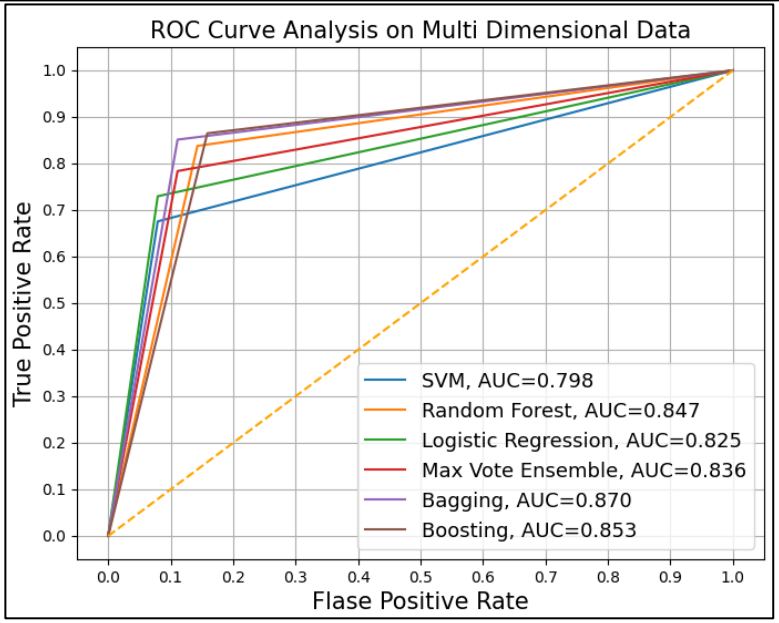


Figure 5. Area Under Curve (AUC) Graph for All Datasets

Table 1. Results Generated by Algorithms Across All Datasets

Data	Model	Precision	Recall	Accuracy	F1	AUC
Two Weeks	SVM	0.87	0.60	0.73	0.71	0.74
	LR	**0.88**	**0.85**	**0.85**	**0.86**	**0.85**
	RF	0.83	0.80	0.80	0.82	0.80
	MVE	0.86	0.83	0.83	0.84.	0.83
	BAG	0.85	0.85	0.84	0.85	0.83
	BST	0.81	0.81	0.80	0.81	0.79
One Month	SVM	0.86	0.57	0.72	0.68	0.72
	LR	**0.86**	**0.74**	**0.80**	**0.80**	**0.80**
	RF	0.80	0.70	0.74	0.75	0.74
	MVE	0.83	0.74	0.78	0.79	0.78
	BAG	0.83	0.77	0.79	0.80	0.79
	BST	0.79	0.81	0.78	0.80	0.77
3200 Tweets	SVM	0.91	0.82	0.86	0.86	0.86
	LR	0.93	0.88	0.90	0.90	0.89
	RF	0.90	0.90	0.90	0.90	0.89
	MVE	0.97	0.93	0.95	0.95	0.95
	BAG	0.97	0.99	0.98	0.98	0.97
	BST	**0.99**	**0.99**	**0.99**	**0.99**	**0.98**
3200 Tweets Multi Variate	SVM	0.91	0.68	0.79	0.78	0.79
	LR	0.92	0.73	0.82	0.81	0.847
	RF	0.87	0.84	0.85	0.86	0.825
	MVE	0.89	0.78	0.83	0.83	0.836
	BAG	**0.90**	**0.85**	**0.87**	**0.88**	**0.87**
	BST	0.86	0.86	0.85	0.86	0.85

Discussion.

Model interpretability is crucial when applying machine learning to sensitive domains like mental health detection. In this research, various models including Logistic Regression, SVM, and ensemble methods were used to classify depression based on tweets across different datasets. Logistic Regression consistently performed well on smaller datasets (Two Weeks and One Month), offering transparency in how features such as word frequency and age correlate with depression status. This interpretability is vital in real-world applications, enabling mental health professionals to understand and trust the model's decisions. However, while ensemble methods like Bagging and Boosting achieved superior performance on larger datasets (3200 tweets), their complexity often results in "black-box" models, making it difficult to trace how specific features contribute to predictions. Additionally, the imbalance in the dataset, with more depression-related tweets than control data, may have influenced the models' performance, potentially skewing results towards detecting depression more effectively than identifying non-depressed users.

In practical applications, these models could be leveraged to develop tools for early detection of depression in social media users, aiding mental health professionals in gaining insights into digital behaviors. However, the imbalance in the data and the lack of interpretability in complex models like Boosting and Bagging present challenges.

Conclusion and Future Work.

This study identified depressed persons using established methods from the literature within this domain. The study utilized six classifiers, including two from the Support Vector Machine (SVM) and Logistic Regression categories, with the remaining four belonging to Ensemble models. Two types of analyses were conducted on the largest dataset of 3200 tweets. Textual analysis and multivariate analysis. Logistic Regression demonstrated superior

performance on the Two-week and One-month datasets. However, the Boosting algorithm excelled when applied to the 3200 Tweets dataset, and the Bagging algorithm performed best on multivariate data of the same size. These findings suggest that high performance in sentiment classification can be achieved by training algorithms on larger datasets. Consequently, trained models can be deployed online to detect depression among social media users or integrated into frameworks used by professionals to gain additional insights into users' digital behaviors. This indicates that machine-learning approaches are effective tools for developing algorithms to identify depression in online environments. Future work could improve these models by incorporating additional features like gender, follower count, number of links, geolocation, and peak activity times, which may provide deeper insights and enhance prediction accuracy. Addressing data imbalance is also crucial to avoid potential biases in model predictions. While this study focused on traditional machine learning algorithms, exploring advanced deep learning techniques, such as the state-of-the-art Bert model, could lead to further performance enhancements and more robust detection capabilities.

Author's Contribution. Fiza Azam. Writing – original draft, methodology, experimentation, visualization, literature review, result reporting. Maha Agro. Writing an original draft, methodology, experimentation, visualization, literature review, and result reporting. Memoona Sami. Methodology, writing –editing and review, validation, result reporting.
Amirita Dewani. Writing –editing and review, validation, result reporting.

Acknowledgment. Nil

Conflict of Interest. The authors declare they have no conflict of interest in publishing this manuscript in IJIST.

Project Details. This research was not part of any project

References.

[1] "Health and Depression", [Online]. Available. https. //www.who.int/health-topics/depression#tab=tab_1

[2] A. M. Helmy, R. Nassar, and N. Ramdan, "Depression detection for twitter users using sentiment analysis in English and Arabic tweets," Artif. Intell. Med., vol. 147, p. 102716, Jan. 2024, doi. 10.1016/J.ARTMED.2023.102716.

[3] B. Chiranjeevi, V. Tejaswini, A. Mahesh, V. Sushmalatha, R. K. Karre, and M. A. Shaik, "Machine learning predictions for finding depression in Twitter," AIP Conf. Proc., vol. 2971, no. 1, Jun. 2024, doi. 10.1063/5.0195849/3296131.

[4] R. Safa, P. Bayat, and L. Moghtader, "Automatic detection of depression symptoms in twitter using multimodal analysis," J. Supercomput., vol. 78, no. 4, pp. 4709–4744, Mar. 2022, doi. 10.1007/S11227-021-04040-8/FIGURES/13.

[5] L. Bendebane, Z. Laboudi, A. Saighi, H. Al-Tarawneh, A. Ouannas, and G. Grassi, "A Multi-Class Deep Learning Approach for Early Detection of Depressive and Anxiety Disorders Using Twitter Data," Algorithms 2023, Vol. 16, Page 543, vol. 16, no. 12, p. 543, Nov. 2023, doi. 10.3390/A16120543.

[6] D. A. Musleh et al., "Twitter Arabic Sentiment Analysis to Detect Depression Using Machine Learning," Comput. Mater. Contin., vol. 71, no. 2, pp. 3463–3477, Dec. 2021, doi. 10.32604/CMC.2022.022508.

[7] V. Tejaswini, K. S. Babu, and B. Sahoo, "Depression Detection from Social Media Text Analysis using Natural Language Processing Techniques and Hybrid Deep Learning Model," ACM Trans. Asian Low-Resource Lang. Inf. Process., vol. 23, no. 1, Jan. 2024, doi. 10.1145/3569580.

[8] Y. W. Abuhasirah, "Data Mining Approaches for Depression Detection on Social Media Twitter Dataset," Stud. Syst. Decis. Control, vol. 503, pp. 199–218, 2024, doi. 10.1007/978-3-031-43490-7_15.

[9] "WHO, 'Depression,' World Health Organization", [Online]. Available. https.

//www.who.int/news-room/fact-sheets/detail/depression

[10] "Depression Statistics Everyone Should Know." Accessed. Sep. 08, 2024. [Online]. Available. https. //www.verywellmind.com/depression-statistics-everyone-should-know-4159056

[11] "Types of Depression. Major, Chronic, Manic, and More Types." Accessed. Sep. 08, 2024. [Online]. Available. https. //www.webmd.com/depression/depression-types#1

[12] D. J. DeNoon, "Symptoms of Depression", [Online]. Available. https. //www.webmd.com/depression/guide/detecting-depression#1

[13] L. S. Radloff, "The CES-D Scale. A Self-Report Depression Scale for Research in the General Population", Accessed. Sep. 08, 2024. [Online]. Available. http. //www.copyright.com/

[14] A. T. Beck, C. H. Ward, M. Mendelson, J. Mock, and J. Erbaugh, "An Inventory for Measuring Depression," Arch. Gen. Psychiatry, vol. 4, no. 6, pp. 561–571, Jun. 1961, doi. 10.1001/ARCHPSYC.1961.01710120031004.

[15] J. S. L. Figuerêdo, A. L. L. M. Maia, and R. T. Calumby, "Early depression detection in social media based on deep learning and underlying emotions," Online Soc. Networks Media, vol. 31, p. 100225, Sep. 2022, doi. 10.1016/J.OSNEM.2022.100225.

[16] W. W. K. Zung, C. B. Richards, and M. J. Short, "Self-Rating Depression Scale in an Outpatient Clinic. Further Validation of the SDS," Arch. Gen. Psychiatry, vol. 13, no. 6, pp. 508–515, Dec. 1965, doi. 10.1001/ARCHPSYC.1965.01730060026004.

[17] R. Salas-Zárate, G. Alor-Hernández, M. D. P. Salas-Zárate, M. A. Paredes-Valverde, M. Bustos-López, and J. L. Sánchez-Cervantes, "Detecting Depression Signs on Social Media. A Systematic Literature Review," Healthc. 2022, Vol. 10, Page 291, vol. 10, no. 2, p. 291, Feb. 2022, doi. 10.3390/HEALTHCARE10020291.

[18] "Identifying Depression on Twitter." Accessed. Sep. 08, 2024. [Online]. Available. https. //www.researchgate.net/publication/305638561_Identifying_Depression_on_Twitter

[19] K. M. Hasib, M. R. Islam, S. Sakib, M. A. Akbar, I. Razzak, and M. S. Alam, "Depression Detection From Social Networks Data Based on Machine Learning and Deep Learning Techniques. An Interrogative Survey," IEEE Trans. Comput. Soc. Syst., vol. 10, no. 4, pp. 1568–1586, Aug. 2023, doi. 10.1109/TCSS.2023.3263128.

[20] A. M. Bucur, A. Cosma, P. Rosso, and L. P. Dinu, "It's Just a Matter of Time. Detecting Depression with Time-Enriched Multimodal Transformers," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 13980 LNCS, pp. 200–215, 2023, doi. 10.1007/978-3-031-28244-7_13.

[21] D. Liu, X. L. Feng, F. Ahmed, M. Shahid, and J. Guo, "Detecting and Measuring Depression on Social Media Using a Machine Learning Approach. Systematic Review.," JMIR Ment. Heal., vol. 9, no. 3, p. e27244, Mar. 2022, doi. 10.2196/27244.

[22] U. Naseem, A. G. Dunn, J. Kim, and M. Khushi, "Early Identification of Depression Severity Levels on Reddit Using Ordinal Classification," WWW 2022 - Proc. ACM Web Conf. 2022, pp. 2563–2572, Apr. 2022, doi. 10.1145/3485447.3512128.

[23] S. S. Korti and S. G. Kanakaraddi, "Depression detection from Twitter posts using NLP and Machine learning techniques," 4th Int. Conf. Emerg. Res. Electron. Comput. Sci. Technol. ICERECT 2022, 2022, doi. 10.1109/ICERECT56837.2022.10059773.

[24] M. Kabir et al., "DEPTWEET. A typology for social media texts to detect depression severities," Comput. Human Behav., vol. 139, p. 107503, Feb. 2023, doi. 10.1016/J.CHB.2022.107503.

[25] W. L. Tey, H. N. Goh, A. H. L. Lim, and C. K. Phang, "Pre- and Post-Depressive Detection using Deep Learning and Textual-based Features," Int. J. Technol., vol. 14, no. 6, pp. 1334–1343, 2023, doi. 10.14716/IJTECH.V14I6.6648.

[26] R. J. Lia, A. B. Siddikk, F. Muntasir, S. S. M. M. Rahman, and N. Jahan, "Depression Detection from Social Media Using Twitter's Tweet," Stud. Comput. Intell., vol. 994, pp. 209–226, 2022, doi. 10.1007/978-3-030-87954-9_9.

[27] A. Shankdhar, R. Mishra, and N. Shukla, "An Application of Deep Learning in Identification of Depression Among Twitter Users," pp. 661–669, 2022, doi. 10.1007/978-981-16-3071-2_54.

[28] K. Sampath and T. Durairaj, "Data set creation and empirical analysis for detecting signs of depression from social media postings," IFIP Adv. Inf. Commun. Technol., vol. 654 IFIP, pp. 136–151, Feb. 2022, doi. 10.1007/978-3-031-16364-7_11.

[29] K. Cornn, "Identifying Depression on Social Media", [Online]. Available. https. //web.stanford.edu/class/archive/cs/cs224n/cs224n.1194/reports/custom/15712307.pdf

[30] B. Kaushik, A. Sharma, A. Chadha, and R. Sharma, "Machine Learning Model for Sentiment Analysis on Mental Health Issues," 2023 15th Int. Conf. Comput. Autom. Eng. ICCAE 2023, pp. 21–25, 2023, doi. 10.1109/ICCAE56788.2023.10111148.

[31] J. D. Akinyemi, A. O. J. Ibitoye, C. T. Oyewale, and O. F. W. Onifade, "Cyberbullying Detection and Classification in Social Media Texts Using Machine Learning Techniques," Lect. Notes Data Eng. Commun. Technol., vol. 181, pp. 440–449, 2023, doi. 10.1007/978-3-031-36118-0_40.

[32] N. Firoz, O. G. Beresteneva, A. S. Vladimirovich, M. S. Tahsin, and F. Tafannum, "Automated Text-based Depression Detection using Hybrid ConvLSTM and Bi-LSTM Model," Proc. 3rd Int. Conf. Artif. Intell. Smart Energy, ICAIS 2023, pp. 734–740, 2023, doi. 10.1109/ICAIS56108.2023.10073683.

[33] A. Amanat et al., "Deep Learning for Depression Detection from Textual Data," Electron. 2022, Vol. 11, Page 676, vol. 11, no. 5, p. 676, Feb. 2022, doi. 10.3390/ELECTRONICS11050676.

[34] U. of P. Philadelphia, "LEXHUB - Lexicon Tool," Lexhub.org, 2022, [Online]. Available. https. //lexhub.org/wlt/lexica.html

[35] D. R. Jyothi Prasanth, J. Dhalia Sweetlin, and S. Sruthi, "Exploring Human Emotions for Depression Detection from Twitter Data by Reducing Misclassification Rate," pp. 127–135, 2022, doi. 10.1007/978-981-16-3802-2_10.

[36] H. Zogan, I. Razzak, X. Wang, S. Jameel, and G. Xu, "Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media," World Wide Web, vol. 25, no. 1, pp. 281–304, Jan. 2022, doi. 10.1007/S11280-021-00992-2/FIGURES/9.

[37] F. Azam, M. Agro, M. Sami, M. H. Abro, and A. Dewani, "Identifying Depression among Twitter Users using Sentiment Analysis," 2021 Int. Conf. Artif. Intell. ICAI 2021, pp. 44–49, Apr. 2021, doi. 10.1109/ICAI52203.2021.9445271.

[38] B. Nurfadhila and A. S. Girsang, "Identifying Indication of Depression of Twitter User in Indonesia Using Text Mining," Int. J. Intell. Syst. Appl. Eng., vol. 11, no. 2, pp. 523–530, Feb. 2023, Accessed. Sep. 08, 2024. [Online]. Available. https. //ijisae.org/index.php/IJISAE/article/view/2663