Highest Accuracy Based Automated Depression Prediction Using Natural Language Processing

S. V. Tharun[1], G. Saranya[1], T. Tamilvizhi[2], and R. Surendran[3](✉)

[1] Department of Computer Science and Engineering, Amrita School of Computing, Amrita Vishwa Vidyapeetham, Chennai, India
g_saranya@ch.amrita.edu

[2] Department of Computer Engineering, Panimalar Engineering College, Chennai, India

[3] Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai, India
surendran.phd.it@gmail.com

Abstract. In today's fast-paced society, psychological health issues such as anxiety, depression, and stress have become prevalent among the general population. Researchers have explored the use of machine learning algorithms to predict the likelihood of depression in individuals. As datasets related to depression become more abundant and machine learning technology advances, there is an opportunity to develop intelligent systems capable of identifying symptoms of depression in written material. By applying natural language processing and machine learning algorithms to analyze written text, such as social media posts, emails, and chat messages, researchers can potentially identify patterns and linguistic cues associated with depression. These patterns may include changes in word usage, tone, and sentiment. The dataset consists of text-based questions on this information channel. At present, machine learning techniques are highly effective for analyzing data and identifying problems. Researchers have conducted comparisons of the accuracy achieved by different machine learning algorithms using the complete set of attributes as well as a subset of selected attributes. In summary, while the potential for AI to aid in mental health diagnosis and treatment is exciting, it's important to proceed with care and consideration for the complexities of the field and the needs of patients.

Keywords: Machine Learning · Natural Language processing · Data Visualization · Artificial Intelligence · Decision Tree Making · Logistic Regression

1 Introduction

Depression is a significant contributor to global disability, affecting millions of people worldwide. While older adults may have a lower incidence of depression, late-life depression (LLD), also known as geriatric depression, still poses a significant public health concern due to its association with increased morbidity, suicide risk, cognitive and physical impairments, and self-neglect. As the global population continues to age,

S. Kadry and R. Prasath (Eds.): MIKE 2023, LNAI 13924, pp. 95–104, 2023.
https://doi.org/10.1007/978-3-031-44084-7_10

it becomes increasingly important to identify and treat LLD. Although the definition of Late Life Depression (LLD) can vary, it typically pertains to depression experienced by individuals who are aged 60 or older, and can encompass both early and late onset episodes. Like depression in younger individuals, LLD can manifest in various ways, with symptoms ranging from subtle changes in mood to major depression as defined in the Diagnostic and Statistical Manual of Mental Disorders (DSM-5). It is important to approach preliminary results with caution and acknowledge that further research is necessary to better integrate artificial intelligence (AI) in mental health research with clinical practice.

Depression is a prevalent mental health condition that can affect anyone. It is characterized by a persistent feeling of sadness, hopelessness, and low energy, often leading to a loss of interest in activities, difficulty sleeping, and changes in appetite. The causes of depression can be complex, with contributing factors ranging from genetics to environment and psychology. While depression can be a debilitating condition, it is treatable with the help of mental health professionals, and treatment options can include medication, psychotherapy, or a combination of both. Seeking help early is essential in preventing the condition from becoming more severe, as untreated depression can lead to serious consequences, including suicide. Depression can have significant impacts on daily functioning, including work, school, and relationships. Some individuals may experience only one episode of depression, while others may have recurrent episodes throughout their lives. It is important to remember that depression is a legitimate medical condition that should be taken seriously, and with proper treatment and support, many individuals with depression can achieve a better quality of life.

NLP can be used to analyze various aspects of text, such as sentiment, tone, and word choice, to identify patterns that may indicate the presence of depression. The project aims to develop a model that can accurately predict depression in individuals using their written text, such as social media posts or chat messages. The machine learning algorithms to predict the stress level during this quarantine with the following Methodology such as Acquisition of data, Pre-processing the data, Splitting the Training and Testing Data, Classification of data and Analysis on the performance of the model.

2 Related Works

Meera Sharma, Sonok Mahapatra, and Adeethyia Shankar employed a Classification and Regression Tree (CART) model, a type of decision tree frequently utilized in machine learning for information mapping. Although CART models have various applications, they used it for binary classification of their data. One issue highlighted in their study is that CART models may not be readily trainable in other languages due to insufficient datasets in those languages. To address the limitations in the dataset, future studies could consider utilizing more current and extensive samples from various industries [1]. Usman Ahmed et al., developed a method based on the standard PHQ-9 questionnaire to extract nine distinct behavior types that align with the Diagnostic and Statistical Manual of Mental Disorders 5 (DSM-V) for measuring depression symptoms from authored patient text. Their method effectively handles uncertainty and interprets the embedding working. One limitation of their approach is that constructing rules requires prior knowledge and expertise in the system.

In the future, research could focus on improving the accuracy of detecting triggering rules and identifying key words to help psychiatrists make accurate notes and diagnoses [2]. Konda Vaishnavi et al., compared the accuracy of five machine learning techniques, including Logistic Regression, Classifier, Random Forest, Decision Tree Classifier, K-NN and Stacking, in identifying mental health issues based on various criteria. Their results indicated that these classifiers were more accurate than other classifiers, with ROC areas between 0.8 and 0.9. However, a limitation of their study was the use of a minimal dataset. Future research could address this limitation by utilizing a larger dataset to improve accuracy [3]. Theodoros Iliou et al., assessed the effectiveness of seven classification algorithms, which comprised the Nearest-neighbor classifier, J48, Random Forest, Multilayer Perceptron, Support Vector Machine, JRIP, and FURIA, on both the raw and preprocessed datasets. One limitation of these data preprocessing methods and classification algorithms is that they may face difficulties in the training phase if the dataset contains a high amount of irrelevant or redundant information, as well as noisy or unreliable data [4]. The proposed approach involves utilizing the strengths of a commonly-used classifier called "XGBoost" to achieve precise classification of data into four categories of mental disorders: Schizophrenia, Autism, OCD, and PTSD. The results of the experiments demonstrate that this approach is highly effective in accurately identifying various types of mental illness [5].

The approach consists of five modules, namely data acquisition, data pre-processing, training and testing data splitting, data classification, and analysis of model performance [6]. The primary benefit of this approach is its ability to precisely identify stress levels by utilizing the dataset obtained during the quarantine period. However, the limitations and issues of this method may become apparent when dealing with raw data and additional test cases. Further research may yield better results by exploring raw data and including more test cases. In the existing work, brain signals are used to find whether the person is in depression or not but it is a complex process since it uses wave signals. It is a physical device with complex mechanism and so, machine learning is not implemented. It is difficult to use in Telemedicine.

3 Proposed Work

The proposed model aims to develop a machine learning model with natural language processing capabilities that can accurately classify whether an individual is suffering from depression or not. Depression is a significant and prevalent issue in society, and managing it is challenging. Consequently, individuals are at a higher risk of experiencing mental distress. The machine learning approach is well-suited to address such complex tasks, as manual analysis of such data can be time-consuming [7]. One potential application of the model is to aid mental health professionals in identifying individuals who may be susceptible to depression and offering them suitable care and assistance. To achieve this goal, the project will entail collecting a dataset of textual samples from both depressed and non-depressed individuals, as well as creating and validating NLP algorithms for analyzing the text [8]. Machine learning can be employed to classify depression by leveraging past data to identify patterns, and enhance the model's accuracy by adjusting its parameters.

The model has potential applications in aiding mental health professionals to identify people who are at risk of developing depression and to provide them with suitable care and support. To accomplish this, the project will necessitate obtaining a collection of textual samples from both depressed and non-depressed individuals, as well as creating and validating NLP algorithms to analyze the text. Various algorithms can be tested, and the most effective model, with the highest accuracy, can be employed for classification purposes. Figure 1 depicts the entity relationship diagram. Figure 2 depicts the operations performed by the system and the user.

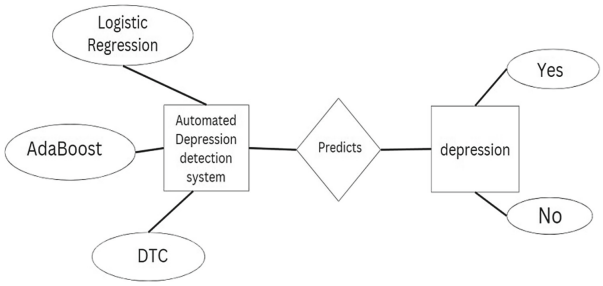

Fig. 1. ER diagram

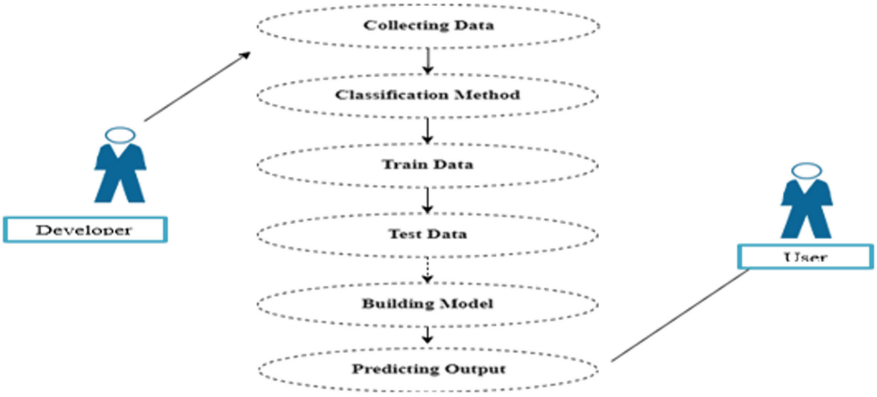

Fig. 2. Use Case diagram

Figure 3 depicts the sequence of activities happening during the prediction of depression when the user clicks the predict option in the website. Figure 4 depicts the Activity during the prediction of depression.

Figure 5 depicts the relationship between the users, interface and the model developed using the processes included as attributes. Both the training and testing data is used to develop the model.

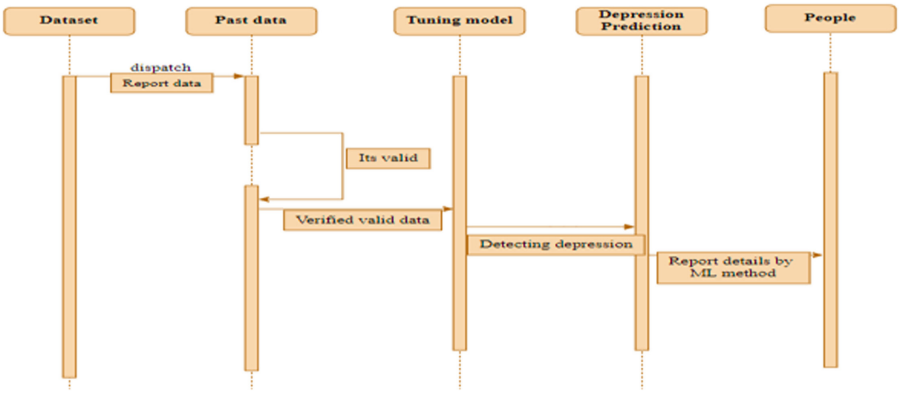

Fig. 3. Sequence diagram

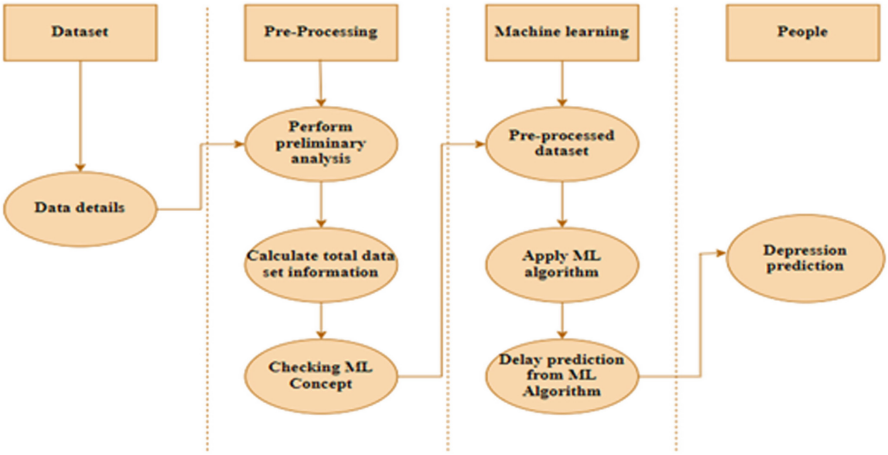

Fig. 4. Activity diagram

3.1 Steps Followed by Decision Tree Classifier

Decision Trees are a form of Supervised Machine Learning that involve splitting data based on certain parameters. This technique involves decision nodes and leaves as the two main entities that comprise a tree. Decision Trees have several real-life analogies and have significantly influenced machine learning, covering both classification and regression [9]. The decision tree classifier works by iteratively dividing the data into smaller subsets based on the most relevant features, until a specified stopping condition is met.

The algorithm selects the best feature to split the data at each node by maximizing the difference in impurity between the parent node and the child nodes. It continues until the leaf nodes are pure or the maximum depth of the tree is reached. The resulting tree can then be used to predict the target variable for new data by following the decision path from the root to a leaf node.

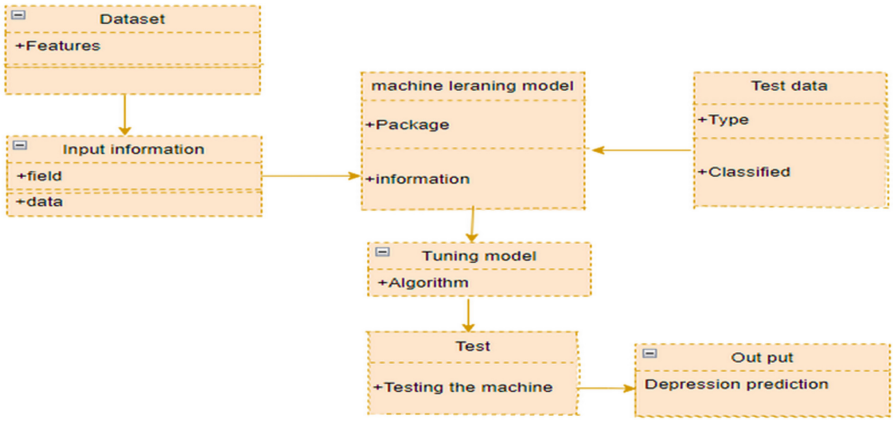

Fig. 5. Class diagram

3.2 Steps Followed in CatBoost

CatBoost is a machine learning algorithm that uses gradient boosting to achieve high accuracy in predictions for numerical, categorical, and textual data. It can be used for classification and regression tasks and is effective in handling noisy data [10]. CatBoost natively handles categorical variables, offers faster training times than other boosting algorithms, and includes advanced techniques like ordered boosting and model introspection for better performance and interpretability. Its capabilities make it a popular choice in various fields, such as marketing, finance, and natural language processing.

3.3 Step Followed in Logistic Regression

Logistic regression is a well-known algorithm in supervised machine learning that is mainly applied to classification tasks. Its primary goal is to estimate the probability of a binary target variable, where the target variable represents either a success/yes (denoted by 1) or a failure/no (denoted by 0). Logistic regression models are created to predict the likelihood of $P(Y = 1)$ based on the independent variable, X. This algorithm is effective and straightforward, and is frequently utilized in tasks such as spam detection, cancer detection, and diabetes prediction [11].

Data Preprocessing

Numpy and Pandas are commonly used Python libraries for data preparation and preprocessing. Numpy is used for mathematical operations and handling multi-dimensional arrays, while Pandas is used for data manipulation and analysis. Together, they provide a powerful toolkit for cleaning, transforming, and restructuring data before it is used for machine learning or other data analysis tasks [12]. To ensure that data can be efficiently processed by the algorithm, it is necessary to carry out data preprocessing, which involves transforming data into a format that is easy to comprehend and manipulate. In Python, the following import statements are commonly used for pre-processing data:

```
import pandas as pd
```

```
import numpy as np
```

Data Visualization

Data visualization is the presentation of data in a pictorial or graphical format. It helps to identify trends, and relationships in the data. There are numerous libraries available in Python that enable the creation of visualizations, including but not limited to Matplotlib, Seaborn, Plotly, and Bokeh [13]. There are various libraries available that offer an extensive collection of functions and tools to create diverse types of graphs, charts, and plots such as line plots, scatter plots, bar graphs, histograms, and heat maps.These visualizations can be customized with different colors, labels, titles, and annotations to make them more informative and appealing to the audience in Fig. 6 and Fig. 7.

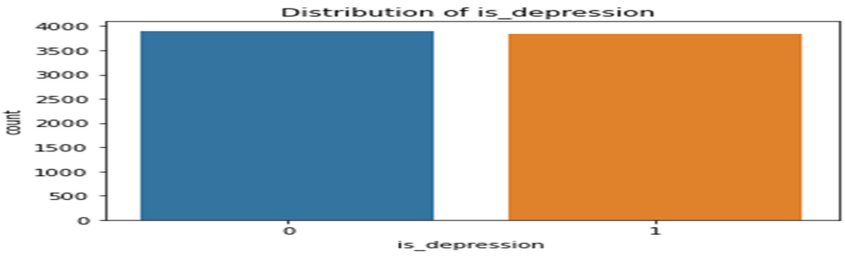

Fig. 6. Distribution Graph

Fig. 7. Word Cloud Image

Algorithm Implementation

In order to effectively compare and evaluate the performance of various ML algorithms, it is essential to establish a standardized test harness. By developing such a test harness in Python, one can gain a template for comparing multiple machine learning models, and can add additional algorithms as needed [14, 15]. Since different machine learning models will exhibit unique performance characteristics, a test harness can be an effective tool for identifying the most suitable algorithm for a given problem [16, 17].

Deployment

In this module, the machine learning model that has been trained is typically saved in a

pickle file format (.pkl file). This file can then be used for deployment in order to create a more user-friendly interface and provide accurate predictions for Depression Prediction with help of Website in Fig. 8. In Fig. 9. Input as Text for the prediction of depression using NLP.

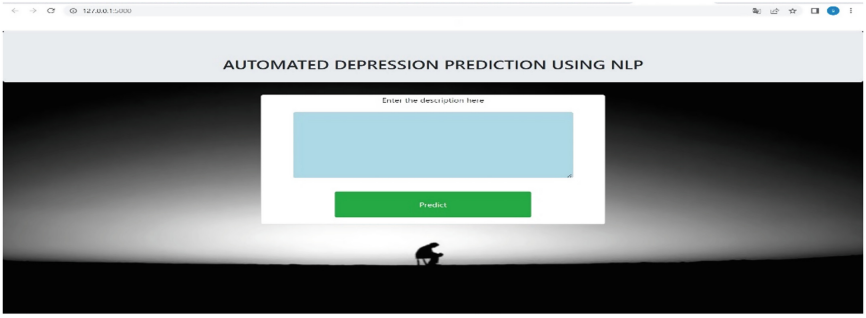

Fig. 8. Website

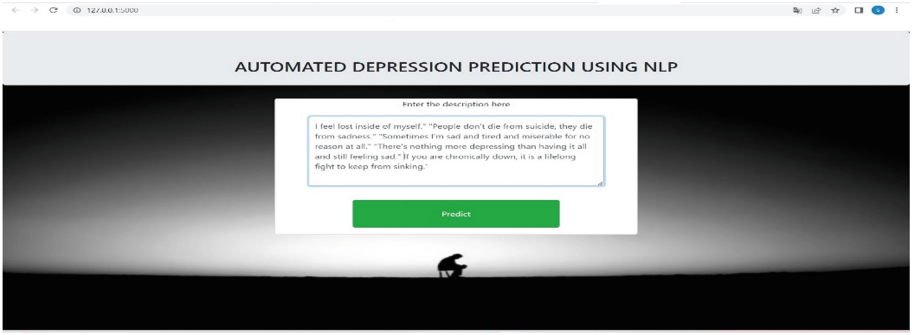

Fig. 9. Input as Text

The result of the proposed work is displayed in Fig. 10 as Predicted output (for depressed person) and Fig. 11. as Predicted output (for not depressed person).

In the proposed system, machine learning will be implemented for more accurate results. It utilized NLP based depression classification was predicted. This system can be implemented in telemedicine, Social Media Platforms to helps in analysing people's mental health. It is deployed in a web browser, so that retains the privacy of the patients. Brain signals are not required in this proposed system. It is a time efficient process.

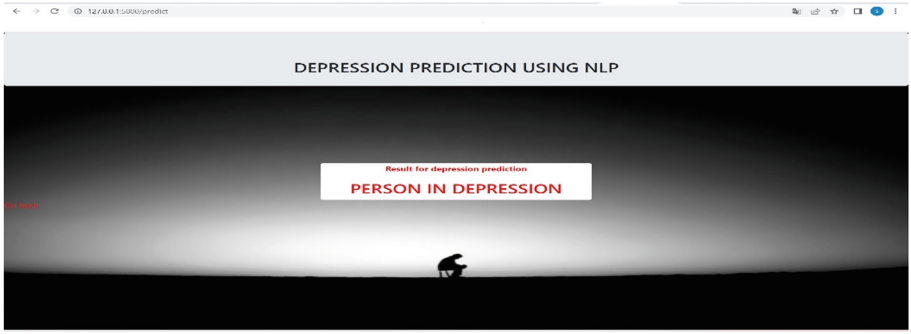

Fig. 10. Predicted output (for depressed person)

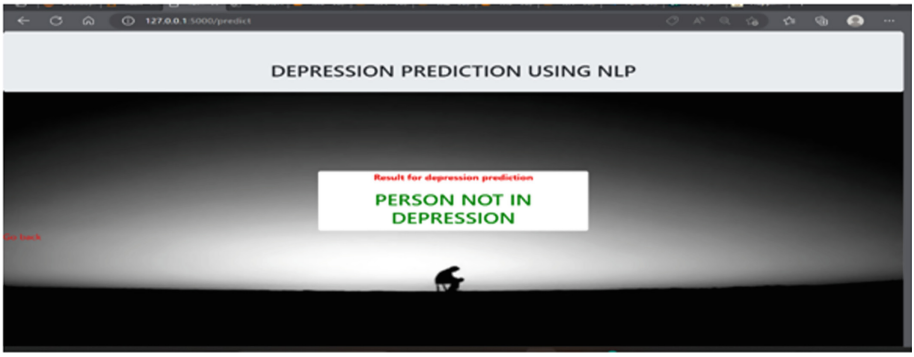

Fig. 11. Predicted output (for not depressed person)

4 Conclusions

The highest accuracy was achieved on a public test set using a high-accuracy scoring algorithm, which was then used in the application to detect depression in individuals. To enhance this system, future work could involve implementing it in multiple programming languages, deploying it on a cloud platform for better accessibility, incorporating more data to improve accuracy, and adding voice input functionality. These enhancements can provide an even more robust and effective tool for detecting depression in individuals.

References

1. Meera, S., Sonok, M., Adeethyia, S.: Predicting the utilization of mental health treatment with various machine learning algorithms. WSEAS Trans on Computers **19** (2019)
2. Usman, A., Jerry, C.L., Gautam, S.: Fuzzy explainable attention-based deep active learning on mental-health data. IEEE International Conference, pp. 6654–4407 (2019)
3. Konda, V., Nikhitha, K., Ashwath, R., Subba, R.N.V.: Predicting mental health illness using machine learning algorithm. Journal of Physics: Conference Series (2021)
4. Theodoros, I., Georgia, K., Mandani, N., Christina, L.: ILIOU Machine Learning Preprocessing Method for Depression Type Prediction, 257–263 (2019)

5. Kamal, M., et al.: Predicting Mental Illness using Social Media Posts and Comments **11** (2021). https://doi.org/10.14569/IJACSA.2020.0111271
6. Thanarajan, T., Alotaibi, Y., Rajendran, S., Nagappan, K.: Improved wolf swarm optimization with deep-learning-based movement analysis and self-regulated human activity recognition. AIMS Mathematics **8**, 12520–12539 (2023)
7. Anishfathim, B., Sreenithi, B., Trisha, S., Swathi, J., Sindhu, P.M.: The Impact of Mental Health due to Covid 19 – A Mental Health Detector Using Machine Learning. Second International Conference on Artificial Intelligence and Smart Energy (ICAIS), **147** (2022)
8. Cho, H.K.: Twitter Depression Data Set Tweets Scraped from Twitter, Depressed and Non-Depressed (2021). Available online: https://www.kaggle.com/hyunkic/twitter-depression-dataset, accessed on 15 January 2022
9. Reya, P.R., Suchitra, S., Gopal, K.S.: The BMI and mental illness nexus: a machine learning approach. International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE) (2020)
10. Piyush, K., et al.: A machine learning implementation for mental health care. 11th International Conference on Cloud Computing, Data Science & Engineering (2021)
11. Soumya Raj, K., Anagha Raj, M., Amulya, N.: Level of stress and coping strategies among institutionalised and non-institutionalised elderly. Indian J. Public Health **11**(03), 637 (2020)
12. Vaibhav, J., Dhruv, C., Piyush, G., Dinesh, K.V.: Depression and impaired mental health analysis from social media platforms using predictive modelling techniques. Fourth International Conference on I-SMAC (2020)
13. Amanat, A., et al.: Deep learning for depression detection from textual data. Electronics **11** (2022). https://doi.org/10.3390/electronics11050676
14. Surendran, R., Karthika, R., Jayalakshmi, B.: Implementation of dynamic scanner to protect the documents from ransomware using machine learning algorithms. In: 2021 International Conference on Computing, Electronics & Communications Engineering (iCCECE), Southend, United Kingdom, pp. 65–70. IEEE (2021)
15. Raymond, C., Gregorius, S.B., Sandeep, D., Fabian, C.: A textual-based featuring approach for depression detection using machine learning classifiers and social media texts. Computers in Biology and Medicine **135** (2021). https://doi.org/10.1016/j.compbiomed.2021.104499
16. Firoz, N., Beresteneva, O.G., Vladimirovich, A.S., Tahsin, M.S., Tafannum, F.: Automated Text-based Depression Detection using Hybrid ConvLSTM and Bi-LSTM Model. In: 2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS), pp. 734–740. IEEE (2023)
17. Ramya, G.R., Bagavathi Sivakumar, P.: An incremental learning temporal influence model for identifying topical influencers on Twitter dataset. Soc. Netw. Anal. Min. **11**, 27 (2021). https://doi.org/10.1007/s13278-021-00732-4