PREDICTION OF DEPRESSION IN SOCIAL NETWORK POSTS USING MACHINE LEARNING ALGORITHMS

B.Vithusa, MAC Akmal Jahan

Department of Computer Science, Faculty of Applied Sciences
South Eastern University of Sri Lanka

ABSTRACT: *Depression is a serious mental disorder and its extreme or worst condition can lead to suicidal action. The number people who suffer from depression is drastically increasing day by day, particularly in teenagers who express it explicitly or keep it invisible which means the depressive feeling is hidden in deep down of their mind. Some of them manages to acknowledge it and some of them even do not know that they are in a depressed mindset. However, this feeling can be emitted in social media pool if the candidate has a habit of posting every event and situation on social networks. Depression silently kills may teenagers and their friends are unknown about it. Since many people maintain social network as an open diary and share everything related to their state of mind, the network users can have the possibility to know the partial scene of the user's situation. If there is a system that can measure the level of depression from users' continuous posts for a certain period of time and give an alert or pop-up notification to friends and family (followers), then we can save many young lives from this tragedy, Therefore, the objective of this work is to build a model by utilizing users continuous posts for a certain period of time. For this, we have investigated machine learning algorithms such as Naïve Bayes, Random Forest, Linear Regression, Support Vector Machine to select a best one with highest accuracy and Support Vector Machine performed better with highest classification performance for the prediction.*

Keywords: Depression, Machine learning, Social Networks, Natural Language Processing, Support Vector Machine.

1. INTRODUCTION

Depression has different kind of effects in our life. Certain symptoms a person expresses can help to predict the status of depression level. Feeling helpless, sad, emptiness or hopelessness, guilty, inferiority complex, angry outbursts, frustration, loss of interests are such psychological symptoms that a person can express explicitly. The depression at its extreme or worst condition even can lead for suicidal ideations (Nafiz et al., 2019). According to the depression, some people express it explicitly, some keeps it invisible which means the depressive feeling is hidden in deep down of their mind. Some of them manage to acknowledge it and some don't know that they are actually facing it.

Nowadays, everyone has a habit of expressing day to day events and situations they experience in their life on social networks. Whether it can be happiness, desire of life or sadness. The social network platform has been becoming as a diary as people share majority of the thing happening in their life and the thoughts of their mind as well. A large group of followers read their posts without knowing the mental status of the user. Now, if we have a pool of social network posts from a person and there is a mechanism to predict the stress or depression level of a person from his/her social network status and posts, then a certain number of people from that network can get aware of the real scenario of a user post by a user alert or any means of pop-up notifications. Therefore, the goal of this work is to predict the depression level of a person from his/her posts using natural langue processing (NLP) techniques and machine learning algorithms. Our proposed works out to select a better algorithm with highest classification accuracy in order to detect depression level by analyzing the data gained from the Social Network sites' users.

2. LITERATURE REVIEW

Depression is a mental health disorder, and the root cause is that something happens in our mind depending on various occasions. For example, sudden changes in our surroundings can cause a change in the brain's neuro transmitter level cause of some shocks or emotional attacks or may be due to genetic features (K. A. Govindasamy and N. Palanichamy, 2021). Depression comes because of various situations – stress of work or at work place, in personal life, or school, college, or because of having other diseases. Depression is a prior condition for incapability or misbehavior over the world and more than 300 million people are undergoing a depression state and about 810,000 people per year commit suicidal attempt (S. Jain et al., 2019), which is being a reason for the death happens among people who are roughly 15-29-year-old adults (S. R. Kamite and V. B. Kamble, 2020). Depression can be cured either by providing a therapy session or prescribing a medication to a patient based on a consultation.

Early detection can assist people to have a proper medical session on time to make their condition get well. Since the last decade, social networks have been growing more and more and are useful in many areas such as sociology, psychology, etc. Nowadays, social network platforms reflect a person's daily life on different stages at different occasions. They provide a possibility to remodel the early mental depressive disorder intercession services (S. R. Kamite and V. B. Kamble, 2020). They allow health experts to obtain knowledge on what is going into a mind of a person who has replied to a topic in a certain manner. To gain such knowledge, machine learning approach offers some advanced features that can help in analyzing unique patterns concealed in online communication and generating them to disclose the mental state by studying the underlying patterns and performing according to it (Nithin et al., 2021). To find the underlying patterns from the user posts, the processes of data acquisition, pre-processing, feature extraction and clarification are performed in the literature.

For the data acquisition, in the existing literature, the researcher used crowdsourcing to fetch data from a user of twitter with identified depression where they used bag of words approach to quantify the data. In another study, author used NCapture tool to collect data from the Facebook and further they used manual methods such as questionnaire, interview, survey for data collection in a direct way. It is noted that dataset was collected from publicly available databases in Reddit (Rafiqul Islam et al., 2018) as well, and poems, songs, and novels were used in the past.

Pre-processing is one of the data preparation processes where NLP techniques were used to extract features to make the data suitable to train. Clean and filtering approaches were used to do the pre-processing. Numeric, empty texts, mentions, non-ASCII characters, hashtags, stop-words, URLs and punctuations were eliminated during the pre-processing (Nithin et al., 2021).

For the feature extraction, the work in (Geetha et al., 2020) considered psychological process, linguistic process and grammar. In the psychological process, all data were labelled based on the user mood (depressed or not), whereas in linguistic process, linguistic inquiry and word count which is a popular tool was used to count the percentage of sentences, texts that gives different states of mind and shared concerns of user data. In another work (Khan et al., 2020) temporal process also considered where it focused on present, past and the future acts of the user. In another study, sentiment analysis was applied for NLP, where text analysis is used to explore the sentiment of the users. Sentiment analysis provided to each data a polarity - positive, neutral, or negative. There were two approaches for sentiment analysis - rule-based analysis (tokenization, stemming, parsing, lexicons) and automatic and hybrid (train model based on corresponding output) approaches.

In another work, researcher applied machine learning algorithms in order to test a predictive power of the classifiers and get knowledge about the influence of feature collections for classifying depression related user posts/tweets (Nithin et al., 2021; K. A. Govindasamy and N. Palanichamy, 2021). Naïve Bayes, Support Vector Machine (SVM), Linear Regression, Random Forest, Decision Tree were used for the depression classification, and they provided the classified output with the accuracy level with the precision and Recall. Even though, many researchers explored different algorithms for different language datasets, there is still a gap exists that needs to find best algorithm for different types of data scenario from different social media platforms.

In addition to the machine learning algorithms, deep learning approach has also been investigated in sentiment analysis in the recent past (Rustam et al., 2021; Zaidi, 2022). Rustam et al., (2021) recently analysed a comparison of supervised machine learning models using covid 19 tweets sentiments where tweets have been extracted using Tweepy library. TextBlob library was used to extract the sentiment, and they have been categorized into positive, negative, or neutral data set. The finding of this work states that the Long Short-Term Memory (LSTM) architecture of deep learning model explicits lower accuracy compared to the machine learning classifiers. In a similar way, the work in (Zaidi, 2022) proposed tweet classification during covid lockdown period using machine learning classifiers with Support vector machine, random forest, KNN, AdaBoost, decision tree convolutional neural network. In this proposed work with our dataset, we have used bag of word approach for feature extraction and we have investigated a set of machine algorithms such as logistic regression, support vector machine, multinomial Naïve Bayes, random forest, KNN to select a better one that suits the data set.

3. METHODOLOGY

The overall flow of the work is outlined as I Figure 1. It consists of following major processes.

1. Data Collection
2. Data pre-processing
3. Data tokenization
4. Classification

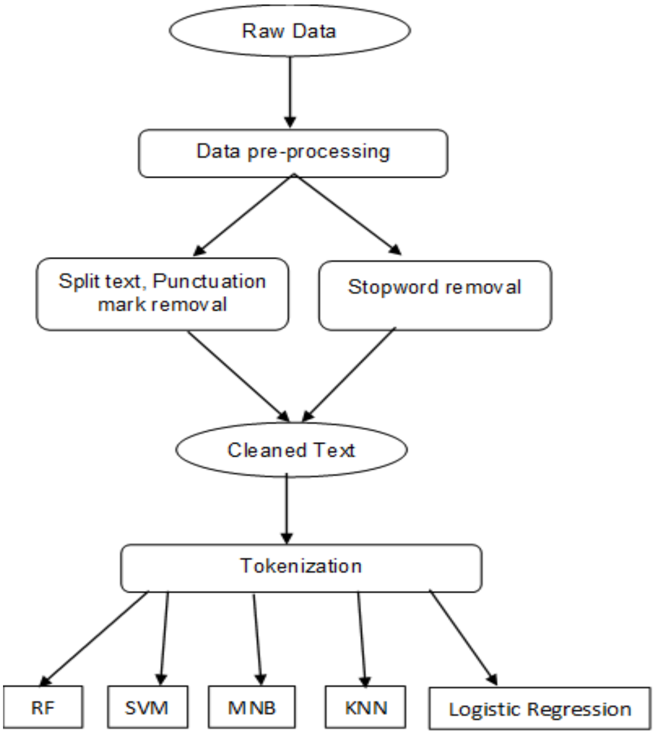


Figure 1: Overall workflow of the process

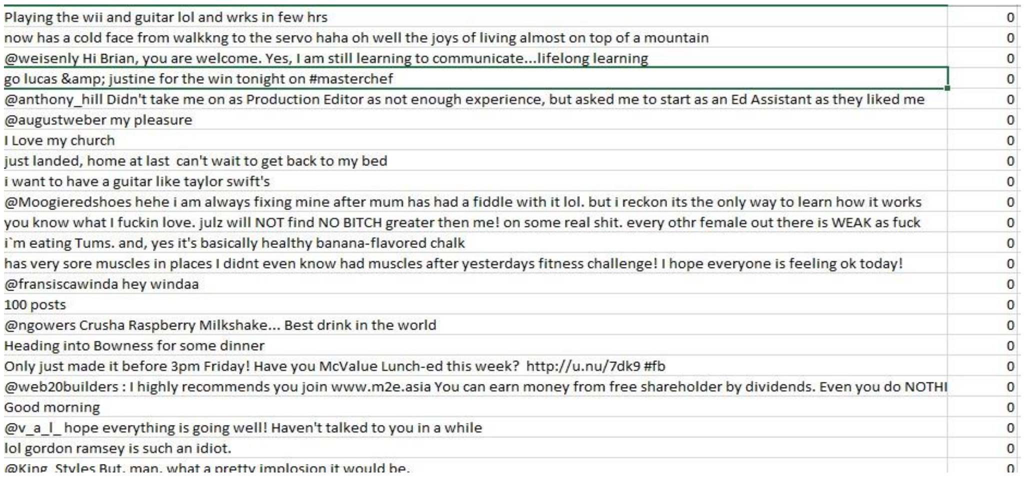
Data Collection

In our proposed model, a set of data were collected from users' social media posts with their permission and some of them were collected from publicly available databases. Then the dataset was processed by using Natural Language Processing modules in order to be trained and tested using machine learning approaches. For our model, we have used two parameters to analyze the data and labelling. One is depressed and the other one is not depressed. Depression is labelled as 1 and the other option is labelled as 0. Table 1 and Figure 2 shows a set of sample social network posts which explicit different emotions of users related to depression.

Table 1: Sample depressed and not depressed expressions

Depressed	Not Depressed
i) Work gets worse and worse each day	My Grandma is making dinner with my mum
ii) I think I'm going to cry at my first day in university next semester.	Just sitting in the garden letting the sun do its job
iii) Having trouble to start today	Looking forward to a mini-break in Isle of Wight with friends this weekend.
iv) I'm so tired but I can't sleep!!	At the library... wow I'm cool

v) Hi twitters! Today i don't feel so good, I'm in a big headache!!	@kevinzahri for that area, it is amazing!



Playing the wii and guitar lol and wrks in few hrs	0
now has a cold face from walkkng to the servo haha oh well the joys of living almost on top of a mountain	0
@weisenly Hi Brian, you are welcome. Yes, I am still learning to communicate...lifelong learning	0
go lucas &amp; justine for the win tonight on #masterchef	0
@anthony_hill Didn't take me on as Production Editor as not enough experience, but asked me to start as an Ed Assistant as they liked me	0
@augustweber my pleasure	0
I Love my church	0
just landed, home at last can't wait to get back to my bed	0
i want to have a guitar like taylor swift's	0
@Moogieredshoes hehe i am always fixing mine after mum has had a fiddle with it lol. but i reckon its the only way to learn how it works	0
you know what I fuckin love. julz will NOT find NO BITCH greater then me! on some real shit. every othr female out there is WEAK as fuck	0
i`m eating Tums. and, yes it's basically healthy banana-flavored chalk	0
has very sore muscles in places I didnt even know had muscles after yesterdays fitness challenge! I hope everyone is feeling ok today!	0
@fransiscawinda hey windaa	0
100 posts	0
@ngowers Crusha Raspberry Milkshake... Best drink in the world	0
Heading into Bowness for some dinner	0
Only just made it before 3pm Friday! Have you McValue Lunch-ed this week? http://u.nu/7dk9 #fb	0
@web20builders : I highly recommends you join www.m2e.asia You can earn money from free shareholder by dividends. Even you do NOTHI	0
Good morning	0
@v_a_l_ hope everything is going well! Haven't talked to you in a while	0
lol gordon ramsey is such an idiot.	0
@King_Styles But, man, what a pretty implosion it would be.	0

Figure 2: Sample user posts with labeling

Data Pre-processing

Pre-processing of data is a mandatory step in natural language processing. It is essential to make the data fits into machine learning model. As a cleaning process, we have performed removal of punctuation marks, regular expression and stop word removals.

Data tokenization

After pre-processing of the text data, the they are tokenized. Tokenization is a process where the sentence is converted into bag of words. Here, we have used *nltk* library and *TfidfVectorizer* for the tokenization.

Classification

There are different types of machine learning based classification algorithms for analyzing the sentiment from a source corpus. The dataset was splitted into training and test data Here, we have selected a set of different algorithms to investigate which one would be better suited for our problem scenario to train our model. The followings algorithms are used in our investigation.

i. Logistic Regression
ii. Support Vector Machine (SVM)
iii. Multinomial Naïve Bayes
iv. Random Forest (RF)
v. Linear support vector
vi. K Nearest Neighbor (KNN)

4. EXPERIMENT AND RESULTS

From the analysis of the algorithms with a classification accuracy, it is observed that the logistic regression has the lowest prediction performance, and SVM and KNN shows highest prediction performance with 96.46% as shown in Table 2 and Figure 3. Further, it is observed that the linear support vector machine is poorly performed than the regular support vector machine.

Table 2: Classifiers with prediction accuracy

Machine Learning Models	Accuracy (%)
Logistic Regression	0.7973
Support Vector Machine	0.9646
Multinomial Naïve Bayes	0.7963
Random Forest	0.8026
Linear Support Vector	0.7293
K Nearest Neighbor	0.9646

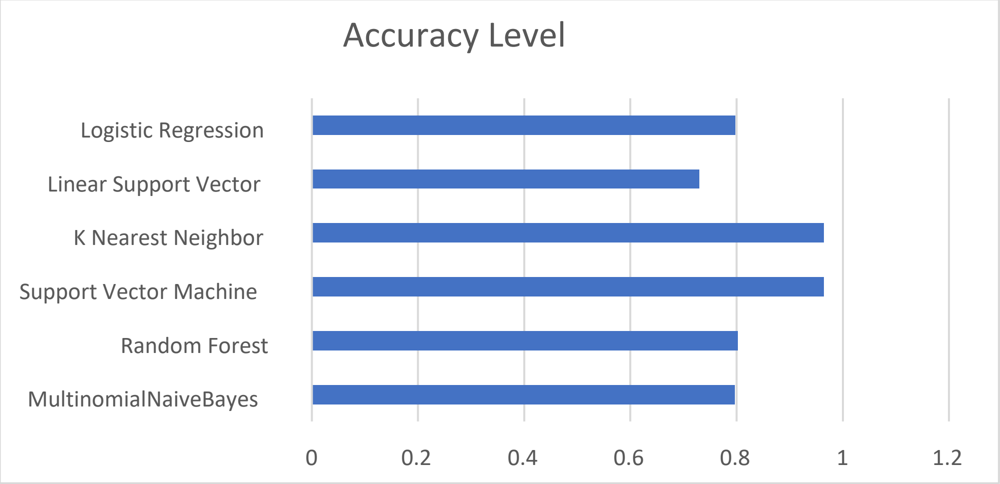


Figure 3: Graphical representation of the results

5. CONCLUSION

Our proposed system predicts and measures the accuracy level of the algorithm for the depression prediction. The support vector Machine model provides 0.96 of highest accuracy level than the other models. By getting best accuracy level, we can choose the suitable machine learning algorithm to process, improve and reach our goal of the proposed system. In future, we plan to consider the depression analysis in Tamil language as well.

REFERENCES

Md. Rafiqul Islam, Muhammad Ashad Kabir, Ashir Ahmed, Abu Raihan M. Kamal, HuaWang and Anwaar Ulhaq, (2018). Depression Detection from Social Network Data using Machine Learning Techniques*, Health Inf Sci Syst*, 6(1).

Nithin P, Owais Ahmed, Aditya Jason Hans, Amaan Faraaz, Prof. Madhusudhan M.V. (2021). Depression Detection Model Based on Sentiment Analysis on Twitter API, *International Journal of New Technology and Research (IJNTR)*, Volume-7, Issue-5, pp. 16-20.

Nafiz Al Asad, Md. Appel Mahmud Pranto, Sadia Afreen, Md. Maynul islam. (2019). Depression Detection by Analyzing Social Media Posts of User, *IEEE International Conference on Information and Communication Technology (SPICSCON),* pp. 28-30.

M. R. H. Khan, U. S. Afroz, A. K. M. Masum, S. Abujar and S. A. Hossain. (2020). Sentiment Analysis from Bengali Depression Dataset using Machine Learning, *11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pp. 1-5.

K. A. Govindasamy and N. Palanichamy. (2021). Depression Detection Using Machine Learning Techniques on Twitter Data, *5th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 960-966.

S. R. Kamite and V. B. Kamble. (2020). Detection of Depression in Social Media via Twitter Using Machine Learning Approach, *International Conference on Smart Innovations in Design, Environment, Management, Planning and Computing (ICSIDEMPC)*, pp. 122-125.

S. Jain, S. P. Narayan, R. K. Dewang, U. Bhartiya, N. Meena and V. Kumar. (2019). A Machine Learning based Depression Analysis and Suicidal Ideation Detection System using Questionnaires and Twitter, *2019 IEEE Students Conference on Engineering and Systems (SCES)*, pp. 1-6.

G. Geetha, G. Saranya, K. Chakrapani, J. G. Ponsam, M. Safa and S. Karpagaselvi. (2020). Early Detection of Depression from Social Media Data Using Machine Learning Algorithms, *2020 International Conference on Power, Energy, Control and Transmission Systems (ICPECTS)*, pp.1-6.

Rustam F, Khalid M, Aslam W, Rupapara V, Mehmood A, Choi GS. (2021). A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis. *PLoS ONE*, 16(2).

Syed Ali Jafar Zaidi, Indranath Chatterjee, and Samir Brahim Belhaouari. (2022). COVID-19 Tweets Classification during Lockdown Period Using Machine Learning Classifiers, *Applied Computational Intelligence and Soft Computing*, 2022(7):1-8.