EasyChair Preprint

№ 4742

Depression Level Detection Using CNN Approach from Tweet Data

Lutfun Nahar, Fahima Alam and Khadiza Sultana Fairose

December 12, 2020

Depression level detection using CNN approach from Tweet data

1st Lutfun Nahar
department of CSE
International Islamic University Chittagong
Chittagong,Bangladesh

2nd Fahima Alam
department of CSE
International Islamic University Chittagong
Chittagong,Bangladesh

3rd Khadiza Sultana Fairose
department of CSE
International Islamic University Chittagong
Chittagong,Bangladesh

***Abstract*—Sentiment analysis is the source of investigation that detect the emotion of people's. It also analyzes the judgments, evaluations, and attitudes from the recorded expression. In natural language processing sentiment analysis is the largest effective field of data mining.The purpose of this investigation is to detect the depressed and non-depressed user from the social media which platform is twitter.After detecting the depressed users from the sentiments,we can counsel them who have the probability of depression to provide proper treatment.This detection has been done by the Machine learning and Deep learning method Support Vector Machine and Convolutional Neural Network.**

***Index Terms*—Depression,Social-media,Machine learning,Deep learning**

I. Introduction

Sentiment analysis is the mechanized procedure of recognizing and removing the abstract data that underlies a text. This can be either a supposition, a judgment, or an inclination about a specific point or subject. The most well-known sort of sentiment analysis is called 'polarity detection' and comprises in characterizing an announcement as 'positive', 'negative' or 'neutral'.The inception of sentiment reasoning perhaps followed to the 1950s,when assumption investigation was basically utilized on composed paper report.Today,be that as it may,sentiment analysis is broadly used to emotional data from content on the Internet, including writings, tweets, online journals, web based life, news stories, surveys, and remarks. This is finished utilizing a wide range of systems, including NLP, insights, and AI techniques. Associations at that point utilize the data mined to distinguish new chances and better objective their message toward their objective socio-economics.Sentiments allude to mentalities, suppositions, and feelings. Various sorts of sentiment analysis utilize various systems and methods to recognize the estimations contained in a particular text . There are two primary sorts of assumption investigation: subjectivity/objectivity identification and feature/aspect-based identification. Sentiment analysis has various uses.It is particularly useful for social media monitoring. Most quite, with the ascent of internet based life sites like Facebook and Twitter. The way people express their perspective, opinions have changed in the age of the internet nowadays, It is essentially done through blogs, journals panel, online conference, product review websites, social media, etc. To explicit their emotions, point of view and share views about their everyday lives, millions of people are using social network sites like Facebook, Twitter, Google Plus, etc. In the pattern of tweets, status updates, blog posts, comments, reviews are the Social media which develop a large number of sentiment data etc.There are so many application method on this platform.Those are: business,politics,public actions and so on. Twitter, one of the biggest social media platforms today.Around 500 million tweets posted by 350 million active users per day.Deep Learning and other Machine Learning (ML) techniques will be applied for helping the process and extracting the valuable data from these large amount of information.This paper focus on performing Sentiment Analysis with Deep Learning and Machine learning on Twitter data to classify it into "positive", "negative" and "neutral" one. Those who needs to expeditiously check the general assessment of tweets focusing on a explicit item, management, association, or another elements could be useful.

II. Purpose of the study

Sentiment analysis is the translation and classification of feelings (positive, negative and neutral) within text data using text analysis strategies.The approximate objects which point out the human mood, affections, and impression of sentiments to identify the underlying emphasis of the interpretation.The purpose of this research is to detect the depressed and non-depressed user from the social media which platform is twitter.After detecting the depressed users from the sentiments,we can refer to counsel them who have the probability of depression to provide proper treatment.depression influences how one can feel, think, act and can interfere with the capacity to work and continue with day by day life.Some symptoms are given below:

1.Absence of enthusiasm for exercises regularly delighted in
2.Changes in weight
3.Changes in rest
4.Feeling of uselessness and guilt
5.The thought of death suicide
6.Absence of vitality

7.Concentration is troubled
8.Anger annoying
Depression has also other reasons which might be happened by family history and genes. To concentrate our exploration will attempt to estimate the nature of separated data.

III. Background Information

To manage issues through information assessment.Data mining is the course toward supervising huge lists to recognize designs and create a relationship. To predict future patterns Data mining gadget used to enable ventures.
Long Ma(), Zhibo Wang, and Yanqing Zhang (Z. Cai et al. (Eds.)2017) proposed to apply text mining from social media sites to extract depression symptoms .They collect their data from Twitter platform. These data were analysed by word frequency,word embedding word clustering. Word embedding follows the neural network language modeling to represent the vector of word.Word2Vec is utilized to consume less time.K-means are used for word clustering to detect the relationships between two words.
Richard J. McNally,Alexandre Heeren,Sanne de Wit and Eiko I. Fried, George Aalbers (2018) presented the Passive social media use (PSMU) for 14 days stress and depression symptoms 7 times daily in 125 students.They utilized the time-series model for the multilevel vector autoregressive.It is also identified the relations between PSMU and specific depression symptoms.In addition with PSMU we incorporated on Active social media use (ASMU) to measure time-spent, in this way the impact of passive versus active use of social media was measured.12-questionnaire were also developed for experience sampling methodology (ESM).
Mandar Deshpande,Vignesh Rao (2017) applied NLP to detect emotions from Twitter feeds concentrating on depression.Class prediction, SVM and Naive-Bayes classifier implemented to identify the accuracy, primary classification matrix,F1-score and the confusion matrix. As the training and test datasets two word-lists were incorporated. Twitter API was involved to collect the tweets as the dataset.
Quan Hu,Ang Li,Fei Heng,Jianpeng Li,Tingshao Zhu(2015) built both classification and prediction from social media via microblogging behavior.The extraction are defined from behavior series.For each behavior series they extracted four features such as: mean,variance,sum and weighted sum. On the other hand,for feature selection Greedy algorithm applied for detecting the sensitive features from depression.
Lang Hea,∗, Cui Caob (2018) built the Deep Convolutional Neural Network (DCNN).By this they manually extracted MRELBP from spectrogram. AVEC2013 AVEC2014 showed that this methodology are more vigorous and effective.They removed the low-level-descriptors (LLD) from the raw audio clips for hand-crafted-features and MRELBP from the spectrograms of sound.By utilizing two different models the deep-learned features were extracted .Firstly, from frame level it removed deep learned audio features and secondly feature representations from spectrogram images learned by this model directly.
Michael M. Tadesse,Hongfei Lin , Bo Xu , Liang Yang(2019) used NLP and machine learning techniques to identify lexicon of terms that are increasingly normal among discouraged records form Reddit users.SVM classifier used with 80% accuracy and 0.80 F1 scores but LIWC,LDA, n-bigram methods were most successfully shown with the Multi Layer Perceptron(MLP)classifier bringing about the top execution at 91%accuracy and 0.93F1scores.

IV. Contribution

1) CNN and SVM have been applied for the proposed method.
2) To extract the depressed sentiments data those who have the probability of depression or anxiety.
3) To extract the depressed sentiments data those who have the probability of depression label and to provide them proper counseling and suggest for proper treatment to come out from this situation. We evaluated the rate of depressed kind of users from the social media.
4) We also categorized the dataset into 2 and 3 category. The two category divided into positive and negative sentiments and the 3 category divided into positive ,negative and neutral sentiments.

V. Deep learning and Machine learning process

A. Deep learning

In an artificial neural network (ANN) a deep neural network is a part of it. It has multiple layers between the input and output layers. The DNN finds to transform the input into the output to correct mathematical manipulation, regardless it is a linear relationship or a non-linear relationship. There is various sorts of neural network in deep learning, for example, convolutional neural systems (CNN), Recurrent Neural Network (RNN), Artificial Neural Network (ANN), and so forth are changing the manner in which we associate with the world.CNN are depicted below:
Convolutional neural network is the declamation of the 3-dimensional convolutional neural system reproduces the straightforward and complex cells of the human mind, including the open fields that people experience through their faculties. CNN is a specific type of artificial neural network. For supervised learning to evaluate data a machine learning unit algorithm is utilized. Image processing, natural language processing and other tasks is enforced by CNN .The CNN is structured differently as compared to a regular neural network. Each layer is composed loads of neurons in regular neural network. Each layer is associated with all neurons in the past layer. In 3-dimensional layers of the convolutional neural network a width, height, and depth are measured for applying. All neurons in a specific layer are not associated with the neurons in the past layer. Rather, a layer is just associated with a little segment of neurons in the past layer.With all the neurons in the successive layer in spite of only a little area of it, the neurons in one layer don't coordinate.Lastly, to a solitary vector of probability scores, formed onward with the depth measurement will be lessened for the last output.CNN

is composed of two major parts.These are feature extraction and classification.In the feature extraction part the convolution and pooling layer is done and the other part is fully connected layer will serve as classifier.

B. Machine learning

In AI the Machine Learning is the examination of computer algorithm that improve consequently through understanding. It is seen as a subset of man-made consciousness. AI calculations construct a logical model dependent on test information, known as "training data", so as to settle on forecasts or choices without being unequivocally customized to do as such.AI counts are utilized in a wide assortment of uses, for example, email sepereting and PC vision, where it is troublesome to create ordinary calculations to develop up required task. In AI strategies two methods are used.These are supervised and unsupervised learning. The supervised study which trains calculations dependent on model input and yield information that is marked by people, and unsupervised study which gives the calculation no notable information to permit it to discover structure inside its instructed documents. Support Vector Machine (SVM) is a labeled AI strategy such could be utilized considering the pair of classification and regression. The principle is the background to isolate the information into two regions which are divided linearly with the most extreme distance. The two-proportion model outlined in SVM finds an optimal edge that lies like a long way from the closest class information focuses as could reasonably be expected. The edges are known as support vectors. SVM acquisition a perfect hyperplane in a high-spatial zone which differs that information along the most extreme edge enclosed by the indicated hyperplane including closest preparing information purposes of each and every region. A kernel capacity is feasibly tested into a higher-spatial zone of the information or data directed toward to compose it directly separable wherever the input information isn't linearly perceptible. The Gaussian radial, polynomial kernel, and sigmoid kernel are non-linear kernels. It utilizes a procedure called the kernel trick to change the data after which principally dependent on these distinctions it uncovers a top of the line limit among the potential output.Simply arranged, it does some exceptionally entangled insights adjustments, at that point makes sense of how to isolate the data-dependent on the labels that have outlined. In this exploration, support vector machine has been applied on the dataset that works utilizing in kernel function.

VI. Methodology

A. Data preprocessing

Data preprocessing is a necessary advance in Machine Learning as the nature of information and the helpful data that can be gotten from it straightforwardly influences the capacity of our model to learn.It is the most important step that devours the more often than not of the exploration.The result of this progression is the last dataset on which the algorithms are applied.We had collected 40 thousands of tweets data from social media.At first we have to clean our data to avoid errors.The cleaned tweets data is added to create a CSV file for training and testing.Then the CSV file has been read and a few data preprocessing steps are performed on it .For preprocessing NLP method has been applied on the extracted data.From this extracted data the sentiments are converted into numerical value.To isolate the post into individual tokens the process of tokenization has been performed.Stop words require to expelled words because there are some words which has no utilization of these words in the preprocessing steps. To expel prevented words from the tweet Nltk library has a lot of stop-words.These words can be utilized as a source of perspective to expel prevent words from the tweet. Punctuations are additionally eliminated which could lead into unstable outcomes if remain avoided.For further preprocessing the arrangement of tweets need to changed over into vector group.The arrangement of names comparing to particular tweet is likewise taken care of into the classifier in the structure a vector.there are two normal methodologies to create word vectors: one-hot encoding and the word embedding.Word2Vec is applied to produce the great nature of word vectors through training the entire dataset.The reason behind of using Word2Vec is to consumes less time.Then the both algorithms CNN and SVM has been applied to predict the accuracy.Data preprocessing has additionally been finished during coding usage in Python Jupyter Notebook to deal with the connection between the factors as there was absolute and numerical information in the dataset.

- Stop word removal
- Punctuation removal
- Tokenization

B. Data Formulation

Through this examination data assortment, data preprocessing, data cleaning, and all the procedures identified with data were not all that simple. In reality, the majority of time has been expended because of data settings. At first, all the information that had been gathered was wrecked up way. The datasets were then arranged appropriately in a manner that is referenced above. The collected information was in a filtered document that has been changed over to CSV records to run the dataset in a python environment. The dataset contains an irregular persistent value which was difficult to process in any model. That is the reason data detailing or formulation for the final evaluation was required in this exploration. All the outputs of executions have appeared here.

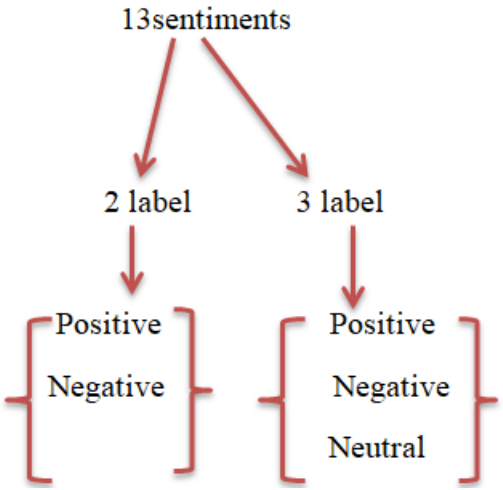

Fig. 1. Data formulation

VII. Result and Analysis

A. Comparison between Depressed and non-depressed users in social media for two and three category

- Two category result

At first, we divided our 40000 thousand data into two labels and find the accuracy with the proposed algorithm. There are 13 sentiments that are converted into positive, negative, and neutral. In the two labels of sentiments, we consider the relief, fun, surprise, love, happiness, neutral, enthusiasm as positive. On the contrary, the boredom, empty, worry, sadness, hate, and angry consider as negative. The number of positive and negative user into two labels are given below:

TABLE I
TWO CATEGORY OF USERS SENTIMENTS

Sentiments	**The number of users**	**The percentage of users**
Positive (non-depressed)	23937	59.84%
Negative (depressed)	16063	40.16%

Performance of the algorithms for two category:

TABLE II
ACCURACY PERCENTAGES OF TWO CATEGORY OF USERS SENTIMENTS

Algorithms	**Accuracy rate**
CNN	98.78%
SVM (linear)	72.05%
SVM (rbf)	72.45%
SVM (sigmoid)	71.25%

CNN gives more accuracy than SVM in this regard.

- Three category result

In the three labels of sentiments,the neutral,boredom empty considerd as neutral label. Then we consider the relief, fun, surprise, love, happiness,enthusiasm as positive. On the contrary,the worry, sadness, hate, and angry consider as negative. The number of positive and negative user into three labels are given below:

TABLE III
THREE CATEGORY OF USERS SENTIMENTS

Sentiments	**The number of users**	**The percentage of users**
Positive (non-depressed)	15299	38.25%
Negative (depressed)	15057	37.64%
Neutral (neutral)	9644	24.11%

Performance of the algorithms for three category:

TABLE IV
ACCURACY PERCENTAGES OF THREE CATEGORY OF USERS SENTIMENTS

Algorithms	**Accuracy rate**
CNN	97%
SVM (linear)	58.025%
SVM (rbf)	58.56%
SVM (sigmoid)	57.091%

After the assessment of the above task, it has been chosen to apply the algorithm which gives more accuracy and from the above plot it has been seen that CNN gives more accuracy. So for the further execution CNN has been applied on the final dataset.

VIII. Discussion and Conclusion

In this research, the proposed model analyze the emotions or sentiment from the social media Twitter platform.We experimented with word embeddings(Word2vec), trained on Twitter data.CNN and SVM algorithms has been utilized to show the accuracy and to know which one is performing better.Utilizing different algorithm made a difference to understand which is appropriate for this framework.All through this research, it is discovered that CNN gives better accuracy than SVM with this big data in two different way.The Convolutional Neural Network gives more accuracy in both way because it is computationally efficient. From the dataset,we read out the sentiments in two category.At first, the two category of sentiment is analyzed where it has determined the percentage of positive, negative sentiment and also to detect few users who have the probability of depression from negative sentiment.In chapter 4,we find the comparison between depressed and non-depressed users from social media twitter platform.A number of users have been found those who are in the situation of depression.The rate of negative user is 40.16% are found which are considered as depression kind of user. So,it is clear in this category,the majority of users feelings have the probability of depression and few people feelings are not misery.
For this two category, CNN has 98.78% of accuracy in two category analysis. SVM(Linear), SVM(rbf), SVM(sigmoid) give the accuracy 72.05%, 72.45%, 71.25% respectively.

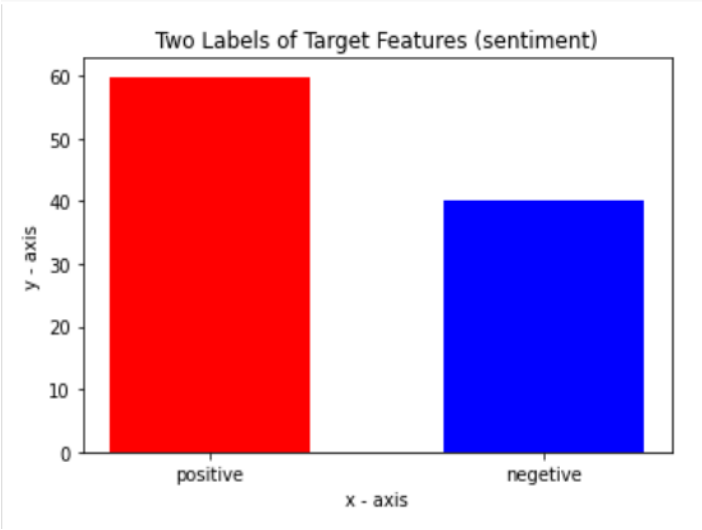

Fig. 2. Two category (positive and negative) of target features (sentiment)

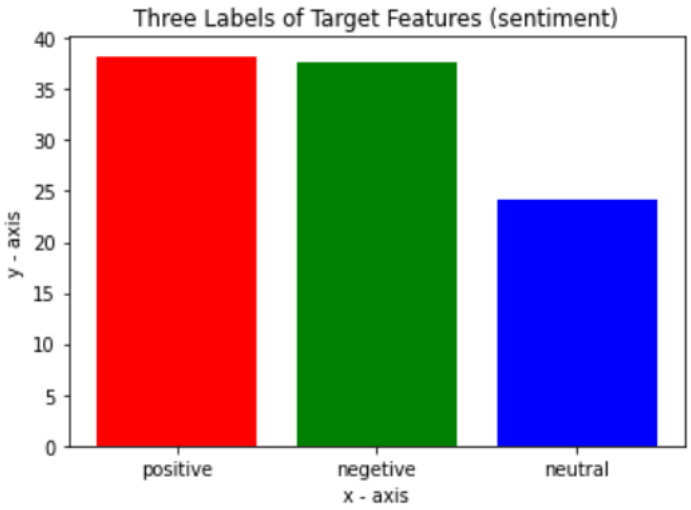

Fig. 4. Three category (positive,negative and neutral) of target features (sentiment)

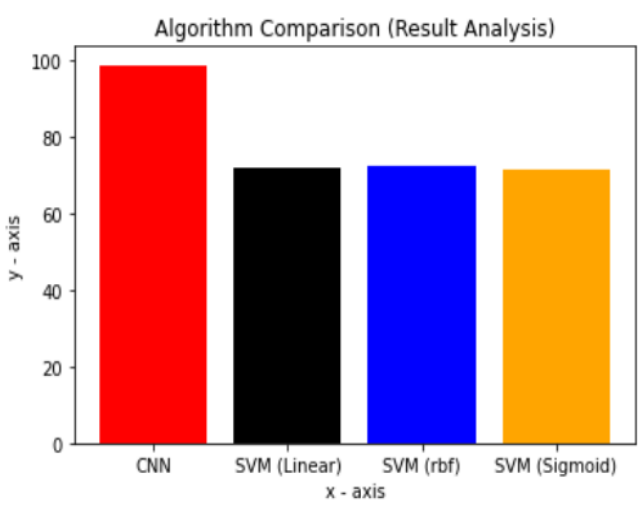

Fig. 3. Two category algorithm accuracy

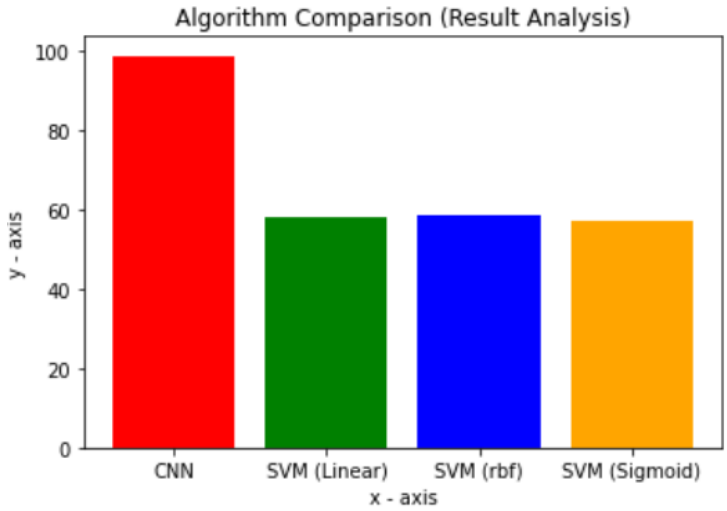

Fig. 5. Three category algorithm accuracy

From the avobe plot,it can say that the percentage of negative user are is sufficient in number at social media.These negative user have the possibility of depression.some symptoms of depression can be figure out from those negative user through their post which they share on twitter platform.

Secondly,the three category of sentiment is analyzed where it has determined the percentage of positive, negative and neutral sentiment.The rate of positive ,negative and neutral are 38.25%,37.64% and 24.11 % respectively.CNN has 97% of accuracy in three category analysis. SVM(Linear), SVM(rbf), SVM(sigmoid) give the accuracy 58.025%,58.56%,57.091% respectively.

Both figure represent the graphical view of positive,negative and neutral sentiment.from the above discussion ,it can be say that the rate of negative user is comparatively high in every perspective. We consider the negative user as depressed kind of user because they might have the possibility of depression symptoms which can be severely depressed or non-severely depressed.These negative user have the bad effect of the depression because they are going through in depression for long or short time period.It also could be possible that they might have the decision of committing suicide.That's why,they need some counselling to come out from this trauma.So that, we can refer them for counselling who have the probability of depression to provide proper treatment.

On the other hand, in every perspective CNN gives mor accuracy than SVM algorithms.The SVM method is learne with the Linear, RBF and sigmoid sector. Though withou exceptions SVM didn't accomplish the better achievement with this dataset.

A. Detection of depressed kind of users

From the dataset,we read out some sentiments of fe users who have probability of depression.The sentiments ar worry,sadness,hate and anger.After read out those sentiment from the dataset the number of worry,sadness,hate and ange are respectively 8459,5165,1323 and 110.These users can b determined as depressed kind of users.

From the table ,the percentage is shown that in social medi worried users are highest in number. That means,form this dataset we can analyze that when a user is in worry they use social media subconsciously and they have high dimensions of depression.Again,when a user is in sad mood the analysis shows that the rate is less than worry disposition.The comparison of hate and angry tendency appears that their range is below than the other two sentiments.Such implies that when a user is in hate and angry attitude they don't use social media frequently or all the time.So,from the analysis,we can estimate that these sentiments have the high probability of depression which can causes many diseases and also something unacceptable can happen like suicide.

TABLE V
PERCENTAGE OF NEGATIVE SENTIMENTS FROM THE DATASET

Depressed kind of Sentiments	percentages
Worry	21.15%
Sadness	12.91%
Hate	3.30%
Angry	0.24%

This figure (fig:6) shows the graphical representation of depressed kind of users.

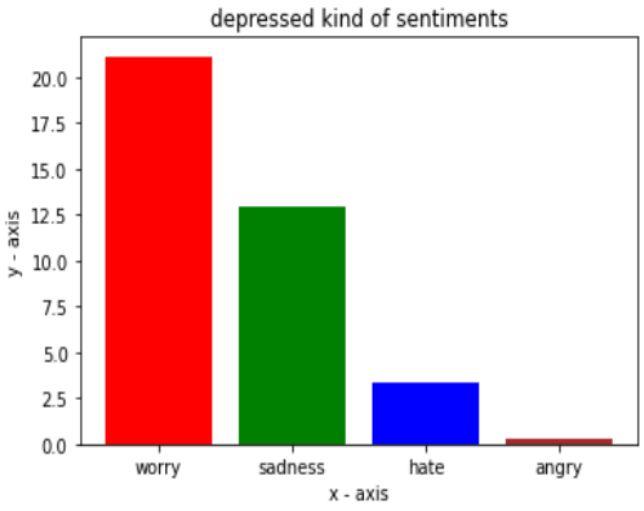

Fig. 6. Depressed kind of sentiments

IX. Future work

As per our research, there is still scope for the improvement in result.Here we have used text type of data for the prediction but by using these algorithms or models different types of dataset can be processed.The dataset can be image or video on sentiment analysis.Here we have read out the depressed kind of sentiment but it can also be done in the further study that the medical symptoms of depressed people would be detected.

References

[1] Durjoy Bapery2, Abu Shamim Mohammad Arif3 Abdul Hasib Uddin1, "Depression Analysis of Bangla Social Media Data using Gated Recurrent Neural Network," in 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT 2019), Bangladesh, 2019, pp. 1-6.

[2] Qi Zhang, Liang Zhu, Xu Zhou, Minlong Peng, and Xuanjing Huang Tao Gui, "Depression Detection on Social Media with Reinforcement Learning," pp. 1-12, 2017.

[3] Vignesh Rao Mandar Deshpande, "Depression Detection using Emotion Artificial Intelligence," in IEEE, 2017), pp. 1-5.

[4] Zhibo Wang1,Yanqing Zhang1 Long Ma1, "Extracting Depression Symptoms from Social Networks and Web Blogs via Text Mining," in ResearchGate, Georgia State University, Atlanta, USA, May 2017, pp. 1-6.

[5] Cui Caob, Lang Hea, "Automated depression analysis using convolutional neural networks from speech," Journal of Biomedical Informatics, vol. 83, pp. 103–111, 29 May 2018.

[6] Alexandre Heeren, Sanne de Wit and Eiko I. Fried,GeorgeAalbers Richard J. McNally, "Social Media and Depression Symptoms: A Network Perspective", 2018," Journal of Experimental Psychology: General, pp. 1-9, September 23 2018.

[7] Gjorgji Strezoski, Gjorgji Madjarov, and Ivica Dimitrovski Dario Stojanovski, "Twitter Sentiment Analysis using Deep Convolutional Neural Network," pp. 1-12.

[8] Chieh-Feng Chiang2 and Arbee L. P. Chen3 Yu Ching Huang1, "Predicting Depression Tendency based on Image , Text and Behavior Data from Instagram," in 8th International Conference on Data Science, Technology and Applications, vol. DOI: 10.5220/0007833600320040, 2019, pp. 32-40.

[9] Pedro M. Sosa, "Twitter Sentiment Analysis using combined LSTM-CNN Models," pp. 1-9, june 7 2017.

[10] S.S. Sonawane Vishal A. Kharde, "Sentiment Analysis of Twitter Data: A Survey of Techniques," International Journal of Computer Applications, vol. 139, pp. 0975 – 8887, 11 April 2016.

[11] Shabib Aftab,Iftikhar Ali Munir Ahmad, "Sentiment Analysis of Tweets using SVM," International Journal of Computer Applications (0975 – 8887), vol. 177, pp. 1-5, 5th November 2017.

[12] Ang Li,Fei Heng, Jianpeng Li,Tingshao Zhu* Quan Hu, "Predicting Depression of Social Media User on Different Observation Windows," in IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2015, pp. 1-4.

[13] PhD Fidel Cacheda1, 2,PhD Diego Fernandez1, 2,PhD Francisco J Novoa1, and 2,PhD Victor Carneiro1, "Early Detection of Depression: Social Network Analysis and Random Forest Techniques," JOURNAL OF MEDICAL INTERNET RESEARCH, vol. 21, no. 6, pp. 1-18, 2019.

[14] HONGFEI LIN,BO XU,LIANG YANG MICHAEL M. TADESSE, "Detection of Depression-Related Posts in Reddit Social Media Forum," IEEEAccess, vol. 7, pp. 1-11, April 16, 2019.

[15] Ang Li,Fei Heng, Jianpeng Li,Tingshao Zhu* Quan Hu, "Predicting Depression of Social Media User on Different Observation Windows," in IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2015, pp. 1-4.