Proceedings of the International Conference on Intelligent Computing and Communication for Smart World (I2C2SW-2018)
14-15 December 2018

# Stress Prediction to reduce Suicidal Rate in the Society based on Social Media using Machine Learning Approach

C.R.Preethi Rajam
*Department of CSE*
*PSNA College of Engineering and Technology*
Dindigul, India.
preethi96sai@gmail.com

N.Uma Maheswari
*Department of CSE*
*PSNA College of Engineering and Technology*
Dindigul, India.

S.Jeyanthi
*Department of CSE*
*PSNA College of Engineering and Technology*
Dindigul, India.

S.K.Somasundaram
*Department of CSE*
*PSNA College of Engineering and Technology*
Dindigul, India

***Abstract*—Psychological Stress is a mental illness which becomes a normal part of life nowadays. The most common human experiences are to exposure the stressful situations like daily annoyances, time-pressured lifestyles, the consequences of overstretched and unexpected events. Stress can motivate and demotivate the person according to their feelings and emotions. It gets difficult when unable to handle on stress and it becomes long-term which can seriously interfere with the health, family life, and job. Due to the overwhelming stress, a person feels that they no longer able to handle it, which leads to suicidal ideation. Nowadays Social media has become a trend and plays a vital role in communication and information sharing worldwide, based on the rapidly growing use of social media and its influence on society, social media service provider offers users a convenient way to create, express, and share their ideas, thoughts, opinions through online comments and tweets. Online communication platforms are increasingly used to express thoughts and analysis of the user's thoughts and opinions which also the important perspective in the business and social environment. By using social media for the analysis of users post to predict the overwhelming stress state of the user in the earlier stage, will reduce suicidal rate. In this paper, we present how to find the stress level of a user by extracting tweet contents like text and images and proposed Optical character recognition, natural language processing techniques and machine learning approach like deep neural network on different social media platforms. This method provides excellent performance and accuracy for real-time data on social media.**

***Keywords—Machine Learning, Natural Language Processing,Deep Neural Network, Social Media***

## I. INTRODUCTION

Psychological stress is an illness that affects people's health. With the rapid pace new advent of lifestyle and due to the economic status, increasingly ubiquitous people are feeling stressed nowadays. Psychological stress is also called clinical depression or mental disorder characterized to exposure the stressful situations like the disruption in the mood, daily annoyances, time-pressured lifestyles, the consequences of overstretched, deliberation and unexpected events. Suicide is the major problem and a leading cause of death due to the psychiatric illness in particular mood disorder. The world population reported the suicide rate in 2018 of the countries worldwide, in which India take 30$^{th}$ rank i.e. about 15.7 suicides per 100k peoples.According to the survey, the suicide rate was increasing in the years 2010 – 2015 about 58,679 people were dead in the software field in India, especially in Tamilnadu about 14 percent (i.e. 8830 people) which is the highest rate of death. In 2016, while about 3 percent of the total premature deaths in India were a result of suicide, the percentage of deaths resulting from self-harm was as high over 9 percent in the five southern states. In the year 2015-2017 at least three documents released by WHO mentioned that the number of suicide per lakh populations in the state at around 28 percent three times of the national average. The World Health Organization document aforementioned 322 million people reside with depression worldwide and nearly half of them boards South East Asian and Western Pacific region, reflective comparatively large populations of India and China. Though depression became more common in individual people's life, overwhelming stress can be very harmful to human health and their personal and official life. Thus, there will be much important has to been given to detect the stress level of the individual before it turns into severe problems. Although the traditional stress detection is examined as standard diagnostic criteria in survey questionnaires or by wearable sensors, these are actually reactive. The traditional method has some limitations like time-costing, labor consuming, hysteretic and computationally expensive.

The rise of social media is changing the society and its environment, technology and in the people's lifestyle, as well as research in healthcare and in the business environment. 'Internet of Everything and Everywhere' made social media and its networks like Twitter, Facebook, WhatsApp, Instagram are become trending nowadays which influence all age group people to use this application to share their feelings and emotions of the day to day life events and mainly to chat with their closed surroundings through the social media. As the continuous use of social media, it is much more possible to identify the psychological state of the user by gathering their data timely. Current social media analysis on posts has been mostly focused on single media data, i.e. either textual orvisual. Suicidal behavior is an important aspect concern to the public health.it is generally considered a preventable death. Suicidal detection and prevention through social media which is much better and more advantage the traditional methods.

## II. RELATED WORK

Nowadays people use more images especially image with quotes (i.e. image with text) in the social media to share their opinion.Detecting and recognition of the text and the characters from the image is a very difficult task because of the blur effect, complex background, low resolution, non-uniform lighting and so no. Generally, the text and character recognition which consists of two methods region-based method and texture based method. Novel weighted Dynamic Time Warping (wDTW) approach was proposed by Vibhor et al in [1] to match the scene images and synthetic images for the text recognition and these images was generated from gradient-based features which is denoted by the lexicon words. Wang et al in [2] every single character is detected by using sliding window procedure and gather the characters to form a word with a limitation of maximum 500 words and has to process. Agrawalet al in [3] implemented a method Maximally Stable Extremal Regions (MSERs) to capture and identify the character candidates. The characters are segmented and it can be detected and MSER size is specified that depends on the range of threshold value to detect the characters. Some of the text recognition can be done by using the Optical Character Recognition method, in with the image was processed and can recognize any character or text in handwritten and printed text also.

Natural language processing (NLP) is a combination of computer science and artificial intelligence. It is the process and analyses large amounts of natural language data with the help of the computer program. NLP which consist of some basic process like Structure extraction, Identify and mark sentence, Language identification, Tokenization, Entity extraction and Phrase extraction, Lemmatization, and Stemming. For sentiment classification Nasukawa et al in [4] implemented NLP with sentiment lexicons. Here natural language processing defines the specific subject using sentiment lexicons expression, and classifies the polarity of the words.

Machine learning is the most important trending technology used for “Decision making” or “Prediction making”. It explores the construction of an algorithm which learns from data and predicts result by strictly following the conditions for higher accuracy. Supervised machine learning which consists of <input, output> pair and the main goal is to learn from the general rule to map the input and output. Here the data are trained to predict the desired result. Machine learning supervised learning approach for classification which plays a major role to predict the stress state of the user in the social network. There are many methods used for classification such as Naive Bayes, neural network, decision tree classification, linear classification, probabilistic classifier etc. By using the machine learning approaches is possible to identify the user behavior and stress level in social media. De Choudhury et al [5] proposed crowd-sourcing techniques, to find the major stress level of users and specifies the use of Twitter social media. This technique uses CES-D (Center for Epidemiologic Studies Depression Scale) for scoring the level of depression either as high or low. Johnson et al in [6] presented a novel approach using NLP and traditional ML classification to monitoring the mental health of Twitter users. Khan et al in [7] proposed hybrid classification-based algorithm along with a Twitter opinion-mining framework using polarity classification algorithm (PCA) consist of the enhanced emoticon classifier (EEC), improved polarity classifier (IPC) and SentiWordNet classifier (SWNC). To analyse and classify the Twitter feed with improved accuracy. O’Dea et al in [8] designed cross-validated machine learning models by using the human codes to identify the difference between the categories and uses tokenization, N-gram (unigram) and bags-of-words. Fuzzy ontology and SWRL technique were proposed by Farman et al in [9] for the sentiment analysis to check the transportation activities and proposed fuzzy ontology-based crawler and semantic knowledge was used in it. Yuan et al in [10] proposed Metropolis-Hastings algorithm in Moodcast to find the emotions in the social network for predicting the emotion which has been modelled effectively for the prediction of each user’s emotion. Lin et al in [11] have introduced a frame with a combination of deep sparse neural network and cross-media microblog data to an automatic stress detection method which is quite feasible and efficient for stress detection. This framework intended to take in the pressure classifications joining the cross-media traits.

In this paper to predict the psychological stress level in social media, Natural Language processing and Deep neural network is mainly focused to improve the performance and accuracy.

## III. PSYCHOLOGICAL STRESS PREDICTION

As the cardinal symptom of psychological stress is severe negative emotions and lack of positive emotions, which can be identified by using the machine learning approaches. The proposed method mainly consists of Natural Language Processing (NLP) and Deep Neural Network (DNN). The main components of the proposed system are Image processing, Preprocessing, Feature extraction, Classification and Connecting to NGO as shown in figure 1. The data was collected from social media like Twitter, Facebook, Instagram, Blogs etc. The image and text data set was used to predict the stress level of the user. The image dataset is collected and processed with OCR which extract the text. And the text tweet content dataset and image extracted text dataset was given as input for preprocessing and feature extraction that uses the NLP. The DNN plays major role that classifies the tweet contents as positive and negative. Finally, the negative tweets and the user details are collected and sent to the NGO for counselling.

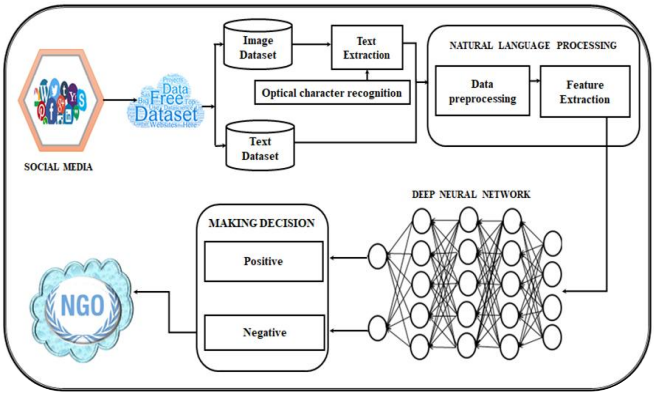


Fig. 1. Architecture of Stress Prediction System

*A. Image processing*

The users of social media to express their feeling, they use to share text, image and video data. In this method, the image data are processed and the text from the image can be extracted with the help of Optical Character Recognition (OCR). The OCR is a method to scans the text sequentially in the images to recognize the separate letters from the text line and it can be grouped. The main processes are Preprocessing, Character recognition and Post-preprocessing. The preprocessing is the process of Noise filtering, Image enhancement and character recognition is important the process which includes Pattern recognition or Adaptive recognition in OCR. The post-processing which makes use of co-occurrence frequencies to correct errors. The preprocessing and post-processing are the important process which are used for lexicon driven OpticalCharacterRecognition to detect the text patterns and with highaccuracy.

*B. Preprocessing*

Natural language processing which consists of some basic preprocessing steps including Structure extraction, Languageidentification,Tokenization, Cleaning, Normalization and so on. Structure extraction is the process of which identifies the fields and blocks of content based on tagging. Tokenization which divides up character streams into tokens. Tokens can be words, numbers, identifiers or punctuation. Cleaning process which removes the stopword and deals with capitalization and characters. Normalization is the process which includes lemmatization and stemming, it uses pattern matching and strip suffixes and reduction to root words accurately.

*C. Feature Extraction*

The features can be used to build a classifier. It can be done with the help of natural language processing approaches. The main approaches used for feature extraction in this approaches are N-gram, POS tagging, and Word2vec. N-gram is computational linguistics and a sequence of tokens for the n-items from a given sequence of text. It is a contiguous syllable, words or multiple words i.e., unigram, bigram, trigram and more which a sequence of the words or parts of speech and it is based on the phrases and expression. Parts of Speech (POS) Tagging is a process of phrases extraction and it simply means labelling each word in a sentence with their appropriate Part of Speech like the noun, verb, adjective, adverb, pronoun, preposition etc. Word2vec is an embedding process that distributed numerical representations of the word features using dimensionality vector. A large text was given as an input and it can produce a vector space. The word2vec which produce the numerical vector values which are given as an input for the neural network for classification.

*D. Deep Neural Network Classification*

Machine learning is the most important trending technology used for 'Decision making' or 'Prediction making'. Neural network classification which gives more accuracy when compared to the traditional machine learning classifications like Linear Regression, Logistic regression, Support Vector Machine, Naive Bayes etc. Deep learning and neural network which is jointly called deep neural network which is used for classification to predict result by strictly following the conditions to produce the results with higher accuracy. The word2vec vector space values are given as an input for the neurons in the deep neural network randomly. Forward propagation is the next step after, here the input weight can be multiplied with the randomly assigned weights of the hidden layer and added the bias value. The activation function is the most important step in the deep neural network, also known as transfer function which converts the input signal to an output signal by passing through this function the first hidden layer was created. The first hidden layer which was given as an input for the next hidden layer and calculation can be done and the out was predicted. If any error occurs it can be identified and changed through back propagation it is the process of learning by mistakes process. The predicted output of the given tweet contents are classified as Positive and Negative values.

*E. Connecting to NGO*

The stressed users are identified by their negative (labelled 0) tweets or comments of the social media users by using the machine learning approaches like natural language process and deep neural network. Then the details of the users like user_id with tweet or comments were collected as a file and it can be sent to the NGO's through cloud service. Then they give counselling to the stressed user to reduce the stress level and the suicidal rate.

## IV. Experimental Result

The dataset consist a mixture of image and text tweet data. The image data is manually collected from the social media and the image was labeled with the user_ id (tweet_id, image). The image dataset is processed with OCR and text can be extracted and combined with the text dataset. The text dataset which consist of training dataset and test dataset which is labeled as (tweet_id, tweet)

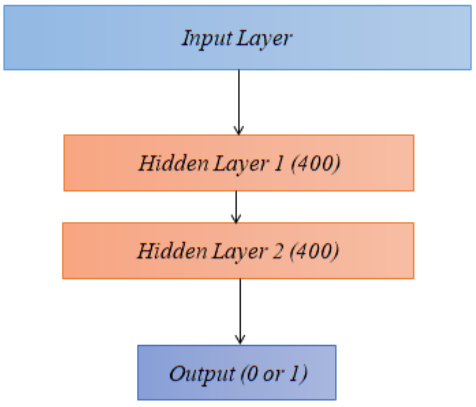


Fig. 2. Deep Neural Network Classification

The tweet contents which consist of tweets with symbols, emojis, URL and so on it can be preprocessed with NLP and features are extracted.

In this proposed DNN classification, more than one hidden layers is used with 400 hidden units shown in Fig. 2. The vector space value is given as input for these hidden

layers and the activation function mainly consist of sigmoid function defined as

$$f(z) = \frac{1}{1+e^{-z}} \qquad (1)$$

The values passes through the sigmoid function and output will be the probability calculation of positive or negative. If the tweet is classified as positive and it can be labeled as 1, if negative it can be labeled as 0.

TABLE I. COMPARISON OF VARIOUS CLASSIFIER WITH DEEP NEURAL NETWORK

| **Classifier** | **Accuracy** | |
|---|---|---|
| | ***Unigram*** | ***Unigram + Bigram*** |
| Support Vector Machine | 79.5 | 80 |
| Naive Bayes Classifier | 78.16 | 79.68 |
| Decision Tree | 68.1 | 68.01 |
| **Deep Neural Network** | **85** | **85.53** |

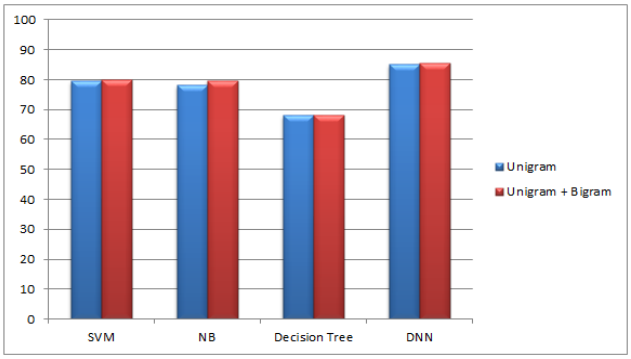


Fig. 3. Comparison of Various Classifier

## V. CONCLUSION

In this paper, we have discussed how to reduce the stress level and the suicidal rate in society by using the machine learning approaches like natural language processing and Deep Neural Network. When compared to the traditional classification algorithms, DNN which gives more accuracy with less response time and ability to handle large data. Here the comparison are made between Support Vector Machine (SVM), Naive Bayes Classifier (NB), Decision Tree, Deep Neural Network (DNN), where the accuracy of SVM is 80, the accuracy of NB is 79.68, the accuracy of Decision Tree is 68.01 and the accuracy of DNN is 85.53. When compared to all other classifier Deep Neural Network (DNN) shows higher accuracy which is shown in TABLE I. and the accuracy comparison show in Fig. 3.

The limitation of the tradition classification algorithms are computationally expensive, high response time and have high complexity and extensive memory required and these are overcome by the DNN classifier.

By using this proposed method the stress rate of the user can be identified by classifying the tweets and comments of the user in social media and the identify the negative tweet and comment as a stressed user with higher accuracy and their details are sent to the NGO's for counselling. Suicide which is not the only solution to solve the problem be strong and ready to face each and every situation in life without any fear and stress which makes life happier and easier.

In future work, we expand the process to include the interaction level as well as the image processing including colours and face recognition which is more effective to identify the stress level of the user and to reduce the stress level and suicidal rate in the society using social media applications.

## REFERENCES

[1] Vibhor Goel, Anand Mishra, Karteek Alahari, C.V. Jawahar, "Whole is Greater than Sum of Parts: Recognizing Scene Text Words," IEEE International Conference on Document Analysis and Recognition, Washington, DC, USA,pp. 398 – 402, October 2013.

[2] Kai Wang, Boris Babenko, Serge Belongie," End-to-end scene text recognition," IEEE International Conference on Computer Vision, Barcelona, Spain, pp.1457 – 1464, January 2012.

[3] Aarushi Agrawal, Prerana Mukherjee, Siddharth Srivastava, Brejesh Lall, "Enhanced Characterness for Text Detection in the Wild," Computer Vision and Pattern Recognition, vol. abs/1712.04927, December 2017.

[4] Tetsuya Nasukawa, Jeonghee Yi , "Sentiment analysis: Capturing favorability using natural language processing," ACM - Proceedings of the 2nd international conference on Knowledge capture, Sanibel Island, FL, USA ,pp. 70-77, October 2003.

[5] Munmun De Choudhury, Michael Gamon, Scott Counts, Eric Horvitz, " Predicting Depression via Social Media," in Proc. 7th ICWSM, USA, pp. 128–37, 2013.

[6] M. Johnson Vioulès, B. Moulahi, J. Azé , S. Bringay, "Detection of suicide-related posts in Twitter data streams," Detection of suicide-related posts in Twitter data streams, Vol. 62 , Issue:1 , 2018.

[7] Farhan HassanKhan, SabaBashir, UsmanQamar, "TOM: Twitter opinion mining framework using hybrid classification scheme," ELSEVIER - Decision Support SystemsVol. 57, pp.245-257, January 2014.

[8] Bridianne O'Dea, Stephen Wan, Philip J.Batterham, Alison L. Calear, Cecile Paris, Helen Christensen, " Detecting suicidality on Twitter," ELSEVIER - Internet Interventions, Vol. 2, Issue 2, pp.183-188, May 2015.

[9] Farman Ali, DaehanKwak, PervezKhan, S.M. RiazulIslam, Kye HyunKim, K.S.Kwak, "Fuzzy Ontologybased Sentiment Analysis of Transportation and City Feature Reviews for Safe Traveling," ELSEVIER - Emerging Technologies, Vol. 77, pp.33-48, April 2017.

[10] Yuan Zhang, Jie Tang, Jimeng Sun, Yiran Chen, Jinghai Rao "Moodcast: Emotion prediction via dynamic continuous factor graph model," IEEE - International Conference on Data Mining, pp. 1193 – 1198, 2010.

[11] Huijie Lin, Jia Jia, Quan Guo, Yuanyuan Xue, Jie Huang, Lianhong Cai, Ling Feng, "Psychological stress detection from cross-media microblog data using deep sparse neural network," IEEE - International Conference on Multimedia and Expo, pp. 1-6, 2014.