Prediction of Suicide Causes in India using Machine Learning

[1]Imran Amin, Sobia Syed

[1,2]*Shaheed Zulfiqar Ali Bhutto Institute of Science and Technology, Karachi Pakistan*

[1]imran.amin@szabist.edu.pk

Abstract—Worldwide, suicide rate is considered one of the most significant issue. With each passing year, the number of suicide is getting increased phenomenally and because of this reason, this research is carried out to predict the causes of suicide in India by using the machine learning algorithms and data mining techniques in order to identify the root causes behind the suicide so that the authorities can take advantage in order to prevent the suicide cases by creating awareness and by rectifying the predicted causes of suicides. According to a research, about 800,000 people commit suicide worldwide every year. Out of these, 135,000 (17%) are residents of India, a nation with 17.5% of world population. In this research, we have analyzed the pattern of suicide cases and predict the causes of future suicides by using machine learning algorithms, the Artificial Neural Network and Support Vector Machine.

***Keywords*—**Machine learning, Algorithms, Data mining, Artificial Neural Network (ANN), Support Vector Machine (SVM).

I. INTRODUCTION

About 800,000 people commit suicide worldwide every year, of these 135,000 (17%) are residents of India, a nation with 17.5% of world population [2]. Between 1987 and 2007, the suicide rate increased from 7.9 to 10.3 per 100,000, with higher suicide rates in southern and eastern states of India [3]. According to the National Crime Records Bureau (NCRB), state of Tamil Nadu, West Bengal, Andhra Pradesh, Maharashtra and Karnataka have registered a consistently higher number of suicidal deaths during the last few years and together accounted for 56.2% of the total suicides reported in the country [4]. Uttar Pradesh, the most populous state (16.5% share of the population) has reported a comparatively lower percentage of suicidal deaths, accounting for only 3.6% of the total suicides reported in this country, but the researcher feels that this is due to the underestimation of suicide cases in this area [1].

This paper studies the prediction of suicide causes in India by using machine learning method and techniques. Although ML has been a part of the computer science field for many decades, it has only recently been applied to clinical psychology. Later, we provide a brief overview to orient readers to what ML is, its advantages over traditional statistical approaches in clinical psychology, and the metrics used to evaluate the performance of ML algorithms.

The purpose of this study is to learn about the trend and changes in suicides rate and to predict the causes of suicide in people of India and to explore and find reasons of increasing ratio of suicides rate and generate a report which can be used in finding solution.

This study has practical and theoretical importance as after the end of research, the outcome of research will be helpful for governmental institutions in India to take action and to find solution by which rate of suicide can be reduce as suicide rate in India is increasing every year.

This research contains dataset of Suicides death in India from year 2001 till 2012 of all the states which is published by National crime record bureau (NCRB) India. The Dataset contain Feature like gender, State, Year, Age group, Total Suicide, Type code and Type.

What are the main causes of increasing suicide deaths and what are its statistics in comparison with other causes?

1) Hypothesis: To analyze the suicidal trend and explanatory association and relationship between suicide rate and economic changes.

2) Limitation of Work: The purpose of this research is to predict the causes of suicide in general irrespective of the age group or gender.

In short, this research is not predicting the causes independently for the every age group or to classify the causes according to the male and female separately.

II. LITERATURE REVIEW

Worldwide, Suicide rate is one of the most important problems. The total number of individuals who committed suicide is increasing with each passing year. It is projected that because of the various causes, around eight hundred thousand individuals expires while attempting suicide [5].

Suicide is considered as a disease and according to the report of WHO (World Health Organization), 17 percent residents of the global suicide sufferers belongs to India [5].

According to the CDC-2015, in the last few years, researchers have focused on recognizing, understanding, curing and impediment of suicidal patterns and behavior. Regardless of all the efforts and studies, the rate of suicide is not decreasing [6].

Majority of the people who attempted suicide does not plan or strategize to attempt a suicide [6].

For that reason, it is very important to make better prediction about the individuals who are expected to take action on their thoughts of attempting suicide.

A researcher projected an integrated framework of machine learning for the prediction of suicide risks. Basically, the proposed structure has three components [7].

1) Temporal characteristic extraction
2) Risk Regulation
3) An ensemble loop for feature selection and ordinal categorization.

Globally, suicide is measured as one the most important issue which leads to the mental health as it is one of the major reason of death. Hence, it is one of the main challenges for the detection and the prevention of suicidal consideration.

For the estimation of suicide rates, the likelihood or probability could be forecasted surrounded by a specified forthcoming era of sentinel measures which are as follows [7]:

1) Low-risk proceedings mean suicide risks are not detected.
2) Moderate-risk measures are self-damage or injuries that does not direct towards the significant consequences.
3) High-risk proceedings are those with major consequences such as deaths.

A research has been published by a researcher which intended to find out the major features that have an effect on the amount of suicide in some particular districts of India and later utilizes those features to estimate the quantity of suicides to be held in future. This suicide estimation can assist or help the authorities in forming leading decisions related to the regions which are affected by high number of suicide [8]. The characteristics in the research represent the fraction of the populace which are distress mainly as a result of suicides [5]. The government of India keeps a record by maintaining a database of the registered cases of suicides for each and every state of India. Database records are made accessible for the public with the intention of analytics of the information present in registered data

Besides, with the amount of suicide cases for every region, the demographical information of that particular state were also considered while developing the estimation model.

There were three basic groups that were considered while developing a model and those categories are educational level, martial stage, and census information of the region.

Researcher applied a Karl Pearson's coefficient of correlation to verify the association of the features and to identify the correlation amongst them. After identifying the strength of association a regression model was applied to for estimating the amount of suicide rate in future.

The conclusive results were significantly important as there as there were nine features which reportedly acquire a significant linear association with the amount of registered suicides.

Estimation model which was developed by utilizing those nine attributes predicted a linear relationship by providing the 99% of estimation accuracy [5].

Another researcher recommends a technique for estimating the suicides. He proposes to utilize the data available for the registered suicides in order to estimate the suicidal behavior amongst individuals. According to him, Sentiment Investigation can play an important role as it is one of the latest experiments developed in machine learning as social networking systems present substantial amount of information and is being gathered and created by the clients/users of the social networking sites. He is in opinion of to extract benefit from the information available at the social networking sites by analyzing the mechanism of the thought procedure which is based upon the opinion, view and the sentiments provided by the user. Social networking platforms are progressively more associated or linked with multiple phenomena like harassment, depression or even suicide cases and because of this it is very important to make an effort to discover the possible sufferers as early as possible so that the

prevention of such incidents like suicides would be achievable [8].

To summarize, author of the research particularly suggest to concentrate on the required terminological sources associated to suicide by means of developing a method for assembling a vocabulary which is correlated with the terminologies of suicide. In this study, Weka Software was utilized which is one of the data mining tools and supports the algorithms based upon the machine learning to investigate and to extract out the meaningful information from the data or the information presented by a Twitter platform.

Thus, as a result an algorithm is proposed along with the mechanism of processing the semantic investigation involving the training data set which were the collection of tweets along with data group established by the tweets on WordNet [8]. Investigational conclusion depicts that the process established on the machine learning technique along with the sentiment investigation can obtain the information of suicidal thoughts or behavior by utilizing the data available at the twitter platform.

Additionally, this study authenticates the helpfulness and efficiency of performance in predicting the suicidal behavior in an individual [8].

III. RESEARCH METHODOLOGY

This section represents the research methodology which has been developed for this particular study. In order to understand the behavior and trend in suicide data, there are two logical methods which deals with our research problem in very effective way. Following are the two approaches for our research work:

1) Descriptive and Statistic Approach: This method is used to find out the pattern of suicides with respect to age group, gender, marital status, social status, education along with the professional occupation.

2) Predictive Approach: In this method, data will be used to generate model to predict future causes of suicide by utilizing the information present in the existing data.

Our research aim is to predict causes of suicide in India based on the existing dataset of India's registered suicide cases which is obtained from NCRB.

Data of the registered suicide cases which is made available for the public by the Indian Government at NRCB website was obtained in order to perform this research.

1) Dataset: In order to investigate, it is important to identify the attributes and characteristics present in a Dataset which contains the information of total suicide in particular state along with other meaningful information which is as follows:

a. *State:*
   This column contains the name of state in India like West Bengal, Andhra Pradesh etc. The total number of unique states which are present in dataset is 35.
b. *Year:*
   This dataset contain information from 2001 - 2012.
c. *Gender:*
   The value in this column is male and female.
d. *Age Group:*
   There are different age group in dataset which are from 0-14 to 60+.
e. *Total Number of suicides:*
   This column contain the sum of total number of suicide in particular state according to its gender, age, and state.
f. *Type/Cause:*
   This column tell us about the reason of attempting suicide like illness, Family Problems, Bankruptcy, unemployment etc.
g. *Marital Status:*
   This column represents the information whether the person who committed suicide was married, unmarried, was a divorcee or a widow.
h. *Professional Occupation:*
   This particular field represents the information whether the victim was a student, employer, house wife or an unemployed person.
i. *Educational Level:*
   This field depicts the information regarding the educational background of a victim.

2) Number of records: The total number of records which are present in our dataset is approximately 109200.

Data pre-processing is one of the most important part for improving the accuracy and performance of our model. Data pre-processing is a data mining technique by which we can clean dataset for reducing redundancy and missing values because real world data or raw data is often incomplete, noisy and also contain error. The selection of incorrect data or feature may result in poor result and accuracy for that reason data pre-processing is necessary.

Pre-processing of data involves multiple step by which we can achieve consistent and complete data. The data pre-processing step are given below:

A. Data Cleaning:

The first step of pre-processing a data is to clean the data. Find data which contain missing values and remove those data or put any other value like average of that column etc. the goal of data cleaning is to remove missing , noisy and inconsistent data from dataset.

B. Data Integration:

Adding or mixing of data from different databases are put together. After gathering all data conflicts with in the whole data is resolved.

C. Data Transformation:

Data is transformed by using normalization or aggregation method.

IV. DATA ANALYSIS AND RESULTS

After the pre-processing of data and developing the methodology for this research paper, extracting meaningful information from that is an important part in order to understand the pattern. The intention or the objective for investigating or analyzing data is to comprehend and interpret the findings of the study correctly in order to draw conclusion from the experiment. For this research paper, we analyzed or examined the data in two steps:

1) Firstly, in this part we have used statistical method to analyze the dataset. The suicide rates in India were calculated for each year from 2001 to 2012 and classified it according to age group and gender. Moreover comparison of suicide rate in elderly and youth as well as comparison of males and females was done to evaluate the difference in tendency and pattern of suicide using SPSS program.

2) The second stage of this data analysis is to predict the causes of suicide through SPSS and MATLAB by applying the algorithms of machine learning on the available data set.

A. Statistical Analysis

a. Age and Gender Effect: By looking at the graph, we are able to identify that most of the females who attempted suicide are aged between 15 to 29 years bracket where as most of the men who attempted suicide are between 30 to 44 age bracket. By looking at the graph in figure 1, we can also analyze that the ratio of the male is comparatively higher than the female who attempted suicide.

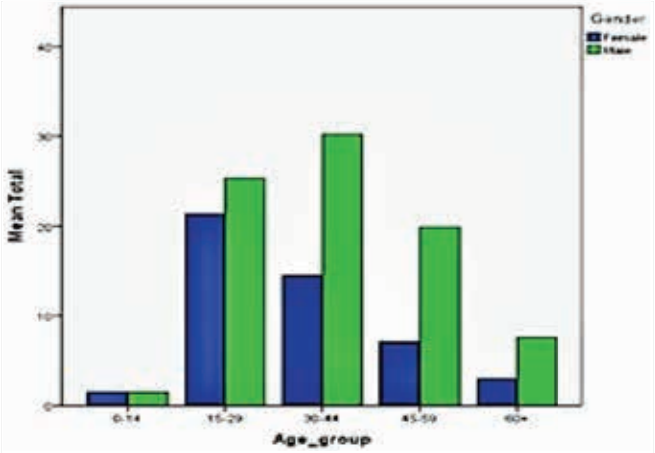

Fig. (1). High suicide rate with respect to age and gender

b. Causes of Suicide: Figure 2, show that the most of the people commit suicide because of the family problems, unemployment, poverty, etc. which leads to the depression and other problems.

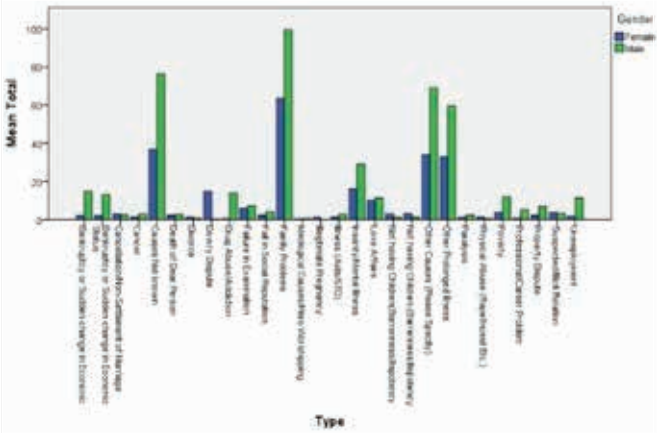

Fig. (2). Causes of suicide

B. Prediction of Causes of Suicide:

For the estimation of causes we have developed a model by using the two very famous algorithms of the machine learning that is the Artificial Neural Networks and the Support Vector Machine.

a. Predicting Model using Artificial Neural Network (ANN): In the field of machine learning Artificial neural network is utilized to extract the information similar to the understanding of human being. This research paper uses the Neural Network based on the multi-layer perception. It consists of the some hidden layers which serve as the intermediate layer between the input and the output layer. In this algorithm, neurons of the similar

layer is not linked or connected but the neurons of the preceding layer are linked with the neurons of the next subsequent layer. For developing the neural network for our research, 70% of the records were divided for the training set of the data whereas the 30% of the data is utilized for the testing and validation. We have given the 10 number of neurons and 6 numbers of inputs age, gender, education, profession, mode of suicide, social status) which gives us the Estimated cause(illness, Love affairs etc.) as an output. Figure 3 show the graphical representation of applied Neural Network Model.

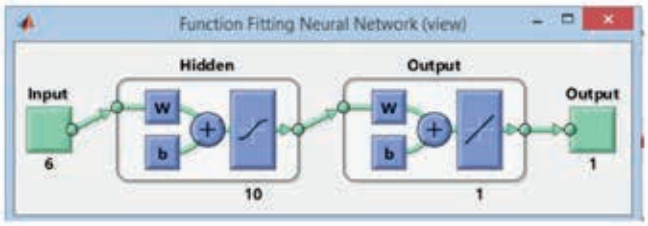

Fig. (3). Applied Neural Network Model

The most important purpose of the neurons at the input level is to distribute the neurons in the middle hidden layer. Input layers Neuron appends the input key xi along with the weights wji of the unconnected association from the input level. The productivity or the output of prototype is Yi and it is equivalent to

$$Yi = f(Pwji\ xi) \quad (1)[9].$$

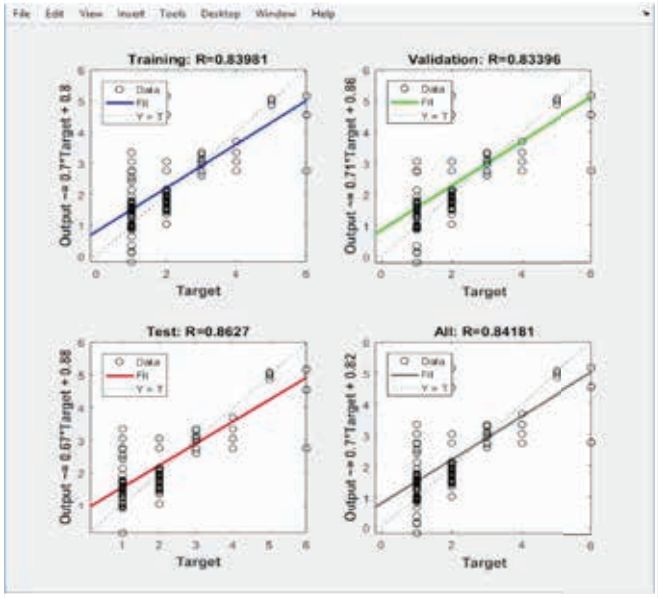

Fig. (4). Regression Analysis on Data

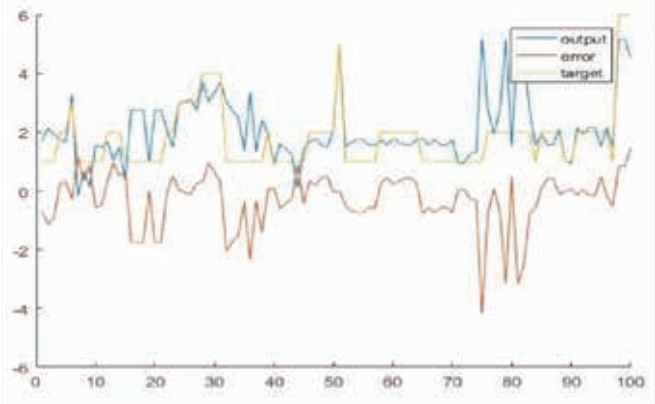

Fig. (5). Predicted Output vs Error vs Actual Output

Similarly figure 5 shows the actual output and simulated output of trained model. We can see that the simulated output line is almost following actual output line. The accuracy of Artificial Neural Network model is 77.5%.

b. Prediction Model using Support Vector machine (SVM): In the machine learning mechanism SVM is another important and considered as the most successful algorithm for the estimation or the prediction of values. Classify generally, two phase process is required to generate the SVM model.

1) Firstly the sample of the experiment data is plotted or recorded on to the significant dimensional area which is very large as compare to the dimension of the original data.

2) Second phase is to discover the ideal hyperplane by means of very large trivial distance in order to categorize data extremely efficient.

For developing the SVM for our research, we have used fivefold cross validation mechanism along with the quadratic kernel function for reviewing the consequences of the investigation result as our goal of this research to predict or estimate the causes of suicide therefore in order to identify the accuracy rate of the prediction we have used this validation.

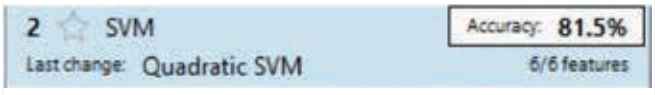

Fig. (6). SVM Accuracy

The figure 6 depicts the 81.5% of accuracy rate of the prediction model for our research so that we can say that out of 10 prediction at least 8 predictions will be accurate by using this model.

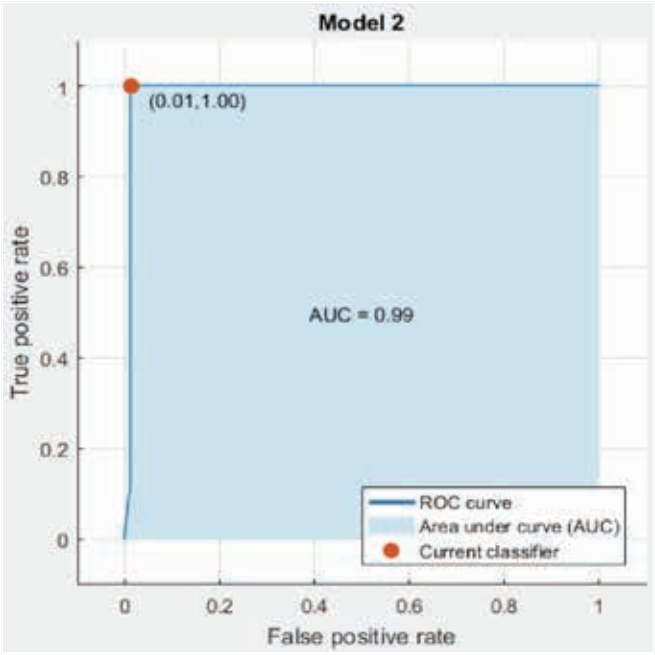

Fig. (7). ROC Curve

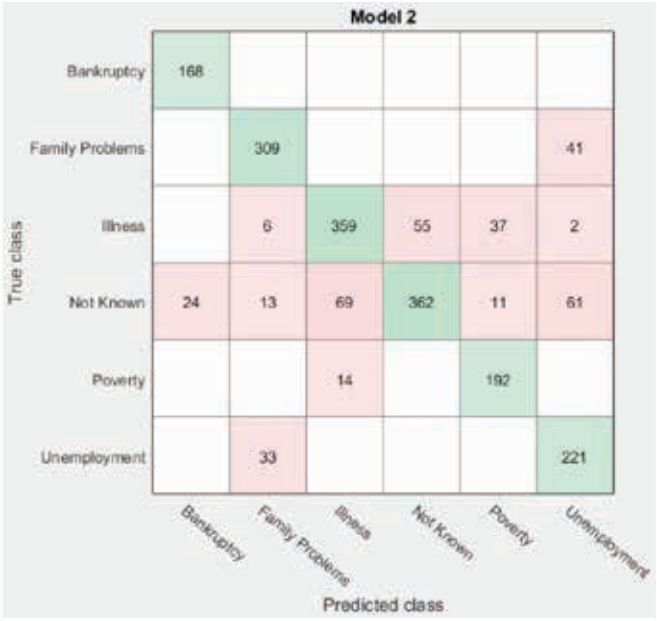

Fig. (8). Confusion Matrix

Figure 7 and 8 of the ROC curve and the confusion matrix shows the implementation of the classification mechanism on the data used for the experiment for which accurate information is known.

V. CONCLUSION

This study is used to analyze the pattern of the registered suicide cases in India. After the analysis of the available data, we can conclude that ratio of the suicide cases for men is comparatively higher than the women. Also, we have identified that the most of the men who attempt or commit suicide belongs to the age group of 30 to 44 whereas the most of the ladies who commit or attempt suicide belongs to the age bracket of 15 to 29. For this research we have developed the two models of Machine learning which are the neural network and SVM for estimating the causes of suicides in future to analyze the accuracy of the both models. We have learned that for this kind of dataset the Neural networks gives the 77.5 % accurate results for the estimation which leads to the 17% of the incorrect predictions whereas SVM model gives the prediction accuracy of 81.5% for the predicting the causes of suicide which makes SVM slightly better than neural network for this particular research.

VI. FUTURE WORK

In future, the applied model of this research can be used in order to predict the causes independently for the every age group and also to classify the causes according to the male and female separately. Also, this research could have also be utilized to predict the amount of suicide in a timely manner.

REFERENCES

[1] S. Kumar, A. K. Verma, S. Bhattacharya, and S. Rathore, "Trends in Rates and Methods of Suicide in India," *Egyptian Journal of Forensic Sciences*, vol. 3, no. 3, pp. 75–80, 2013.
DOI: 10.1016/j.ejfs.2013.04.003

[2] C. G. Walsh, J. D. Ribeiro, and J. C. Franklin, "Predicting Risk of Suicide Attempts Over Time Through Machine Learning," *Clinical Psychological Science*, vol. 5, no. 3, pp: 457-469, 2017.
DOI: 10.1177/2167702617691560

[3] M. Simon, E-S. Chang, P. Zeng, and X. Dong, "Prevalence of Suicidal Ideation, Attempts, and Completed Suicide Rate in Chinese Aging Populations: A Systematic Review," *Archives of Gerontology and Geriatrics*, vol. 57, no. 3, pp. 250–256, 2013.
DOI: 10.1016/j.archger.2013.05.006

[4] V. Arya, A. Page, J. River, G. Armstrong, and P. Mayer, "Trends and Socio-Economic Determinants of Suicide in India: 2001–2013," *Social Psychiatry and Psychiatric Epidemiology*, pp. 1–10, 2017.
DOI: 10.1007/s00127-017-1466-x

[5] M. C. Podlogar, A. R. Gai, M. Schneider, C. R. Hagan, and T. E. Joiner, "Advancing the Prediction and Prevention of Murder-Suicide," *Journal of Aggression, Conflict and Peace Research*, vol. 10, no. 3, pp: 223-234, 2018.
DOI: 10.1108/JACPR-08-2017-0309

[6] J. D. Ribeiro, J. C. Franklin, K. R. Fox, K. H. Bentley, E. M. Kleiman, B. P. Chang, and M. K. Nock, "Self-Injurious Thoughts and Behaviors as Risk Factors for Future Suicide Ideation, Attempts, and Death: a Meta-Analysis of Longitudinal Studies," *Psychological Medicine*, vol. 46, no. 2, pp. 225–236, 2016. DOI: 10.1017/S0033291715001804

[7] J. D. Ribeiro, X. Huang, K. R. Fox, and J. C. Franklin, "Depression and Hopelessness as Risk Factors for Suicide Ideation, Attempts and Death: Meta-Analysis of Longitudinal Studies," *The British Journal of Psychiatry*, vol. 212, no. 5, pp: 279-286, 2018. DOI: 10.1192/bjp.2018.27

[8] M. Birjali, A. Beni-Hssane, and M. Erritali, "Machine Learning and Semantic Sentiment Analysis Based Algorithms for Suicide Sentiment Prediction in Social Networks," *Procedia Computer Science*, vol. 113, pp. 65–72, 2017. DOI: 10.1016/j.procs.2017.08.290

[9] S. Ayat, H. A. Farahani, M. Aghamohamadi, M. Alian, S. Aghamohamadi, and Z. Kazemi, "A Comparison of Artificial Neural Networks Learning Algorithms in Predicting Tendency for Suicide," *Neural Computing and Applications*, vol. 23, no. 5, pp. 1381–1386, 2013. DOI: 10.1007/s00521-012-1086-z