International Journal of Pure and Applied Mathematics
Volume 118 No. 20 2018, 165-171
ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue

ijpam.eu

ANALYSIS OF SURVEY DATA TO PREDICT SUICIDALITY

Kranthi Nuthalapati, RanjithaKorrapati, and Dr. S. Thenmalar

SRM Institute of Science and Technology, Chennai, India
nkranthi28@gmail.com, ranjitha.korrapati@gmail.com, thenmalar.s@ktr.srmuniv.ac.in

Abstract— **Suicides have always been a major concern as it not only affects the person committing suicide but also their family members, employers and the society. A robust method to find suicidal ideation among people is yet to be established. With technology at our finger tips, we can use machine learning techniques to predict suicidality of an individual by studying the data associated with that person. So far, various tweets, blogs, census data and various other forms of numerical data have been analysed. However, our proposed work makes use of a survey with categorical variables. The categorical variables have been analysed and a machine learning model has been developed using logistic regression to build a classifier that classifies the person as suicidal or not. Logistic Regression is one of the best methods to apply on a dataset with categorical values. The machine learning model used has an accuracy of 0.82.**

Key words: Suicides, Machine Learning, Logistic Regression, Categorical Variables, Feature Selection, Economic Behaviour.

I. INTRODUCTION

Behavioral Economics or Behavioral science is the study of human behavior and its implications. For instance, people with anxiety, depression, social fear etc.; tend to be suicidal. Every year around 1 million people die due to suicides worldwide [1]. Therefore this issue requires serious attention and a solid system is required to automate the process of suicide detection leveraging data from various sources like social media, blogs, clinical data of suicidal patients and various surveys and questionnaires [2]. The factors attributing to suicidal tendency can be carefully studied and analysed to come up with a robust system to help find suicidality prior to suicide attempts in people. Various datasets are available with factors like depression, social fear, education status, employment status, marital status etc.; along with suicidality of an individual. Such data sets can be used to come up with a system that finds correlation between various factors and suicidality of a person. For instance, a separated woman is highly likely to commit suicide over a married woman [3].

In this paper we considered a survey from Reddit, a social discussion website that includes questions on various factors like anxiety, depression, sexuality, etc.; most of which are categorical variables. We performed feature selection and extracted important features. Logistic Regression was applied to selected features to obtain results. The paper is organized into the following sections: *Related work* that discusses various methodologies employed so far for suicide analysis, *Survey data set* that gives description of the data set used in our model, *Methodology* which includes the systematic steps followed in building our model, *Experiment and Results* that includes the steps and experimental results and finally *Conclusion* that concludes the paper.

II. RELATED WORK

The following are the categories of suicide analysis:

Social media based suicide detection systems:

As large volumes of data is available on social media like Twitter, Facebook etc.; suicide analysis can be performed by mining huge quantities of text from social media. A twitter based suicide detection system can be made which involves three steps- Data collection by twitter API, Human coding, and Machine Classification. Various machine learning models like SVM, Naive Bayes classifiers etc.; have been applied and SVM was found to have highest accuracy [4]. Social Networking Sites (SNS) like twitter can also be used find geographic location of the person tweeting suicidal content to provide necessary help and support from nearby hospitals, NGOs and relatives [5]. Apart from twitter data, various other SNS like MySpace also have been considered. Comments from MySpace were sorted and potential victims with suicide ideation were recognized [9]. In America, 1 in 8 is prone to problematic internet use. Such a gruelling situation adds the risk to lay

oneself open to various Social Network Mental Disorders (SNMD). Studies point that people suffering from SNMD have a greater risk of suicide ideation after analysing various factors like overuse, depression, social withdrawal etc. with the help of SVM [10].

Suicide analysis using blogs:
Although SNS are quite popular, there have been instances where data from various blogs was studied to find suicide ideation. For example, in China where two million people try to commit suicide and 5% succeed in doing so, the Sina Microblog was studied for suicide ideation among its users using NLP and machine learning techniques [6].

Suicide analysis using census data:
Huge amount of statistical data is provided by government of India [7]. Analysis on this data was performed to find correlation between number of suicides and various other factors like marital status, education status and multiple linear regression models were developed to predict suicidality [3]. Apart from the aforementioned factors, loneliness, child abuse sexual assault and their interaction also have a strong impact on suicide ideation [11],[12].

Suicide analysis using clinical data and questionnaires:
Apart from all these, 30% of the people attempting suicide express their suicide ideation prior to their attempts [13]. It has also been noted that 85% of the patients committing suicides obtain health care or have some kind of medical history [14]. Over a span of three years, the proportion of recurrent suicide was observed to be 19.5 [15]. To study the recurrent patterns and the history of the patients, primary data is obtained via the Electronic Medical Record and analysed using various NLP and supervised learning techniques [13],[16]. Despite the advanced technology and efforts, traditional questionnaires like Questionnaire of Suicide Attitude (QSA), Beck Suicide Ideation Scale (BSS) etc. were used to obtain data for analysis.

III. Survey Data Set

The data set has been taken from Kaggle [8] which has over 500 entries of survey taken from Reddit users. The survey was taken by redditers to elucidate various factors like depression, social fear etc.; the survey had various factors that are significantly important to decide suicidality of an individuals. Some of the questions are as follows:

- What is your Gender?
- How old are you?
- What is your level of income?
- What is your race?
- How would you describe your body/weight?
- How many friends do you have?
- Do you have social anxiety/phobia?
- Are you depressed?
- What kind of help do you want from others?
- Have you attempted suicide?
- Employment Status: Are you currently…?
- What is your job title?
- What is your level of education?
- What have you done to try and improve yourself?

All the important factors after feature selection were taken as independent variables and committing suicide as dependent variable to apply logistic regression on the data set.

IV. Methodology

The following steps (modules) were performed in order to arrive at results:

Data Collection: Data has been collected from the on line data science platform Kaggle, an online platform for data scientists [8].

Data Preprocessing: Data Preprocessing is the process of getting data ready in order to attain best results .Data preprocessing includes various steps like data cleaning, filling empty fields, variable encoding etc.; Data Preprocessing is a very important step as noisy data interferes with the results and may lead to misleading implications. As most of the fields in data set are categorical variables, there is a need to encode the variables before applying a logistic regression model to it. Data was cleaned by filling null records followed by encoding categorical variables. Dummy variables were created for all the fields with categorical variables to convert them into binomial variables (i.e 0 or 1).

Visualization: Various graphs have been plotted to understand relationship between variables in a better way. Stacked bar charts were plotted with various factors like depression (Fig 1), body weight (Fig 2), social fear, employment status, educational status etc.; Data Visualization is the first step to get a bigger picture of the data and understand the important features. From the plots we can understand that factors like existence of depression or over weight have relation with suicidal tendency.

Splitting data into test and train: Data was divided into test and train setsso that train data set can be fitted with logistic regression and further used for prediction of suicidal tendency on test data set. This is done so as to test the prognostic ability of the model on data not used for estimation. 20% of the data was considered for test set so that it will not leave us with too small train data set which might hinder the model to fit and learn the data efficiently.

Feature Extraction and selection: In machine learning, feature is the attribute or property of the case under examination. Feature extraction is a procedure that involves eliminating the redundant data from the dataset. As all the features might not prove useful for our model and may cause over fitting, useful, explanatory and independent feature selection is the key to a noteworthy classification and regression. Choosing correct variables is also crucial for the application of machine learning algorithm. Recursive Feature Elimination (RFE) has been performed to select the best features. Some of the features that have been shortlisted after recursive feature elimination are bodyweight (over weight), depression, particular employment status like (unable to find work) etc.

*Logistic Regression:*Logistic Regression also called logit regression is a regression model which is a very useful model where the dependent variable contains categorical data (eq. 1). The outcome of the model is dichotomous. The logistic regression is used to find the most suitable fitting model which describes the relationship between the independent and dependent variable. As the data set is composed mostly of categorical data, logistic regression was applied to train data with the features selected by recursive feature elimination as independent variable and prior suicidality (0 means no prior suicide attempt and 1 implies suicide attempt).

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta + \beta 1 x 1 + \ldots\ldots\ldots + \beta n x n \quad (1)$$

Where $X=(x_1, x_2, x_3 \ldots x_n)$ and $\log\left(\frac{p(X)}{1-p(X)}\right)$ is the link function [17].

Cross validation: Cross validation is a technique used to analyse the predictive performance of a statistical model. Cross validation is performed to check whether the statistical analysis is accurate or not and also to check for over fitting of data.

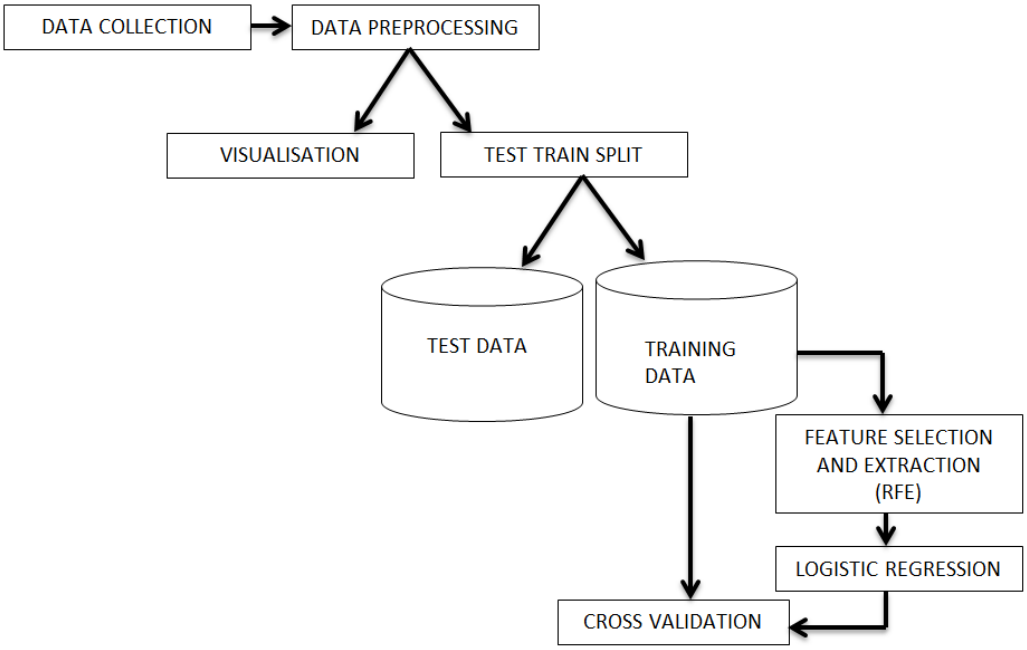

Fig. 1.Architecture Diagram.

V. Experiment and Result

The models used till now were validated using accuracy, precision, recall and Receiver Operating Characteristic. A model was built to predict if a person committed suicide based on various factors taken from survey and validated using accuracy measure (eq. 2).

$$Accuracy = \frac{\text{Number of correct prediction(s)}}{\text{Total number of all cases to be predicted}} \tag{2}$$

The data was processed to avoid noise and split into test and train data to apply logistic regression. Graphs were plotted to understand relationship between suicidality and other factors (fig 2). A greater percentage of depressed people commit suicide to those who are not depressed. Overweight people commit suicides in higher percentages (fig 3). In order to perform analysis, categorical data was converted into binary 0s and 1s and logistic regression was applied. The following factors were among those factors selected using recursive feature elimination:

- gender_Male
- gender_Transgender male
- bodyweight_Overweight
- depressed_Yes
- employment_Out of work and looking for work
- employment_Unable to work
- edu_level_Some high school, no diploma
- edu_level_Trade/technical/vocational training

Upon applying logistic regression, the accuracy of the model on test set was found to be 0.82. In order to ensure validity, cross validation was also performed to get a score of 0.83. ROC curve was also plotted (fig 4).

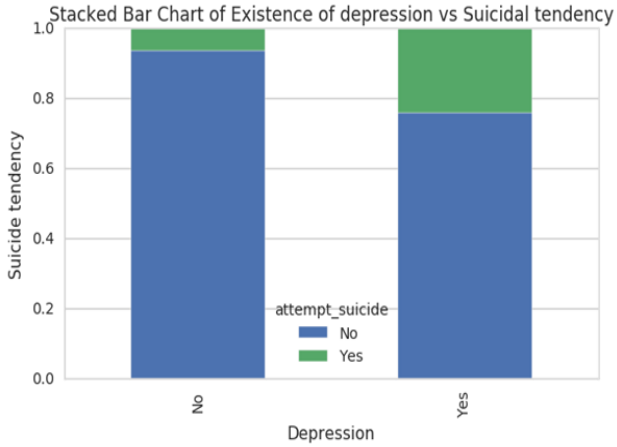

Fig. 2. Stacked Bar Chart of Existence of depression vs. Suicidal tendency.

Fig. 3. Stacked Bar Chart of body weight vs. Suicidal tendency.

Fig. 4. Receiver operating characteristic.

VI. Conclusion

The model suggested in the paper gave highly accurate result(accuracy measure of 0.82) .Based on the features selected using RFE, depressed people, overweight people, and unemployed people etc.; have higher likelihood of attempting suicide. So, proper initiatives have to be taken to curb these issues to suicide attempts. Real time data like medical records of such people have to be collected and further enhancement has to be done in order to develop a robust suicide detection systems. Such systems can be used by police, hospitals, family members and Non-Governmental Organizations to help people with suicidal tendency.

VII. References

1. Grunebaum, M. F.: Suicidology meets "big data", J. Clin. Psychiatry, 76(3), e383-e384 DOI: 10.4088/JCP.14com09421(2015)
2. our paper
3. Priyanka, S.S., Galgali, S., Priya, S.S., Shashank, B.R., Srinivasa, K.G.: Analysis of suicide victim data for the prediction of number of suicides in India, International Conference on Circuits, Controls, Communications and Computing (I4C), pp. 1-5, DOI:10.1109/CIMCA.2016.8053293, Bangalore, India (2016)
4. B. O'Dea *et al.*: Detecting suicidality on Twitter, Internet Interventions volume 2 Issue 2, 183–188 (2015)
5. M. E. Larsen *et al.*: We Feel: Mapping Emotion on Twitter, IEEE Journal of Biomedical and Health Informatics, vol. 19, no. 4, pp. 1246-1252, DOI:10.1109/JBHI.2015.2403839 (2015)
6. X. Huang *et al.*: Detecting Suicidal Ideation in Chinese Microblogs with Psychological Lexicons, IEEE 11th Intl Conf on Ubiquitous Intelligence and Computing and IEEE 11th Intl Conf on Autonomic and Trusted Computing and IEEE 14th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UTC-ATC-ScalCom), pp. 844-849, DOI:10.1109/UIC-ATC-ScalCom.2014.48, Bali, Indonesia (2014)
7. https://data.gov.in/catalog/stateut-wise-distribution-suicides-means-adopted
8. https://www.kaggle.com/kingburrito666/the-demographic-rforeveralone-dataset
9. Cash, S J., Thelwall, M., Peck, S N., Ferrell, J Z., Bridge, A J.: Adolscent Suicide Statements on MySpace, Cyberpsychology, Behavior, and Social Networking, 16(3): 166-174, DOI:10.1089/cyber.2012.0098 (2013)
10. H. H. Shuai*et al.*: A Comprehensive Study on Social Network Mental Disorders Detection via Online Social Media Mining, IEEE Transactions on Knowledge and Data Engineering, vol. PP, no. 99, pp. 1-1, DOI:10.1109/TKDE.2017.2786695 (2017)
11. E. C. Chang *et al.*: Loneliness under assault: Understanding the impact of sexual assault on the relation between loneliness and suicidal risk in college students, Personality and Individual Differences, Volume 72, Pages 155-159 (2015)
12. Sachs-Ericsson, N.J., Stanley, I.H., Sheffler, J.L., Selby, E., Joiner, T.E.: Non-violent and violent forms of childhood abuse in the prediction of suicide attempts: Direct or indirect effects through psychiatric disorders?, Journal of Affective Disorders, Volume 215, Pages 15-22 (2017)
13. Poulin, Chris et al. "Predicting the Risk of Suicide by Analyzing the Text of Clinical Notes." Ed. Vladimir Brusic.*PLoS ONE* 9.1: e85733, DOI:10.1371/journal.pone.0085733 (2014)
14. Ostacher, Michael J. et al: A clinical measure of suicidal ideation, suicidal behavior, and associated symptoms in bipolar disorder: Psychometric properties of the Concise Health Risk Tracking Self-Report (CHRT-SR**)** , Journal of Psychiatric Research, Volume 71, 126-133 (2015)
15. Liu, Y., Sareen, J., Bolton, J.M., Wang, J.L.: Development and validation of a risk prediction algorithm for the recurrence of suicidal ideation among general population with low mood, Journal of Affective Disorders, Volume 193, Pages 11-17 (2016)
16.Ben-Ari, A., Hammond, K.: Text Mining the EMR for Modeling and Predicting Suicidal Behavior among US Veterans of the 1991 Persian Gulf War, 2015 48th Hawaii International Conference on System Sciences (HICSS), pp. 3168-3175, DOI:10.1109/HICSS.2015.382, Kauai, HI, USA (2015)
17. Mathur, P., Khatri, S. K., Sharma, M., Prediction of aviation accidents using logistic regression model, *2017 International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions) (ICTUS)*, Dubai, United Arab Emirates, pp. 725-728, DOI: 10.1109/ICTUS.2017.8286102 (2017)