Proceedings of the International Conference on Electronics and Sustainable Communication Systems (ICESC 2020)
IEEE Xplore Part Number: CFP20V66-ART; ISBN: 978-1-7281-4108-4

Performance Evaluation of Different Machine Learning Techniques using Twitter Data for Identification of Suicidal Intent

Anirudh Ramachandran, Akshara Gadwe, Dishank Poddar, Saurabh Satavalekar, Sunita Sahu (*Authors*)
Computer Department
Vivekanand Education Society's Institute of Technology
Mumbai, India

***Abstract*—Rise of Social media influence has brought about immense changes in the way a person lives their life. Sharing thoughts, ideas and expression over a public platform gives a deep insight into the person's state of mind commonly coined as online behaviour. Research and Evaluation based on online behaviour have been conducted repeatedly. Using machine learning, this online trail of data that a person leaves behind can be used to gain insights on the behaviour and psychological status. In this paper, different machine learning techniques have been used, studied and gauged their effectiveness for suicidal tendency detection to prove that Machine Learning Algorithms like Logistic Regression can correctly identify residing Suicidal Tendency of a Twitter user.**

Keywords—Suicide, Depression, Machine Learning, Natural Language Processing

I. INTRODUCTION

Suicide has always had its position amongst the top 10 causes of death all over the world. It was estimated by the World Health Organisation (WHO) that every year approximately one million people committed suicide, which brings the mortality rate to 16 people per 100,000 or one death every 40 seconds[1]. It is predicted that the suicide rate is going to rise to one every 20 seconds. Younger generations have now replaced elderly males as the group at highest risk in quite a few countries. Mental health disorders (such as depression and substance abuse) are credited with more than 90% of all cases of suicide[2]. Many countries recognise the need and positive impact of Suicide Prevention Strategies, and are working to ensure they are in place. One such prevention strategy is early identification of suicidal ideation or depression among individuals.

The organisation of the paper is as follows. Section II discusses the current scenario. Section III reviews the literature.Section IV gives an overview of Machine Learning. Section V discusses the parameters to Measure Performance. Section VI discusses the different ML Algorithms used by us. Section VII reviews the Discussion and Conclusion.

II. CURRENT SCENARIO

Currently, to detect suicidal tendencies, most psychologists rely on questionnaires and academic interviews [3]. While there are many other issues in understanding the mental state of a patient from their response to the questionnaire, the complete process being voluntary is the main stumbling block. One has to take the initial step, which is something a depressed or suicidal person might hesitate to do because poor mental health is considered to be a personal matter that has to be handled by the individual themselves [4]. So, even if there are many experts available to help mentally ill people, it is not possible for them to actually reach out to the patients unless the patient is willing to volunteer.

Psychologists are looking at web media, to analyze the content put out on social media to gain psychological insight [3]. But, much more psychological research is needed. The challenges associated with this study are numerous because text analysis is difficult. On account of that, no robust system has been built yet that automatically classifies tweets and detects the suicidal ones.

III. LITERATURE REVIEW

Here the various papers published by various researchers regarding suicidal detection using machine learning have been discussed:

In [5], Ryu, S., et. al. performed the analysis on data obtained from Korea National Health and Nutrition Examination Survey (KNHANES). The questions of the survey were used as features to classify the people as suicidal or not. Features which included "stress level in daily life", "EQ-5D : usual activities" etc. Feature elimination using Random forest reduced the number of features from 49 to 35. The model achieved accuracy of 0.781, sensitivity of 0.771, specificity of 0.792. The model with 15 features predicted suicide ideators with an accuracy of 0.821, sensitivity of 0.836, specificity of 0.807.

The advantages of this approach are that machine learning algorithms can successfully classify suicide ideators using general survey data. Also, the physical and mental health data accompanied by Quality of Life data are good indicators to predict Suicide Ideation. However, a simple model such as used here will allow for increased interoperability, reduced number of dimensions make generalisations easy. Further, to make the most of the survey data, suicide ideation instead of suicide attempt was used for prediction. Use of more than one machine learning algorithm is also expected for better performance of the model.

The authors in [6] identified certain tweets, obtained from Twitter API, indicating depression and suicidal tendency using commonly used English phrases and separately stored the

profile username and photo of the respective tweet. With the help of Human intervention the seriousness of the tweet was determined by classifying them into three categories namely , Strongly concerning , concerning and safe to ignore. Scikit Learn Toolkit was used on this human coded data. Two text classification algorithms Support Vector Machine (SVM) and Linear Regression were also tested in [6]. Use of Human coding and Machine Coding together adds one extra degree of confirmation and the rates of agreement amongst the human evaluators made sure that discrepancy was avoided.But, use of limited number of suicicde related terms affected the performance of the classifier. The authors were also unable to differentiate between passive suicide ideation and immediate danger of taking action .

Ji, S., et al. used Reddit and Twitter as the major sources of the data in [3]. In reddit, along with "SuicideWatch" subreddit, other popular subreddits/data was also collected, in order to validate the features extracted from the suicidal ones. Tweets were filtered using keywords and the annotation was done manually. LIWC, TF-IDF and statistics were used for feature extraction. Classification was done by Random Forest, GBDT, LSTM, XGBoost, MLFFNN,etc. They manually annotated the data and did comparison and analysis of suicidal tweets against the ones which were said to be non suicidal.

In [7] the objective of Jashinsky, J., et al. was to check if the suicide rate generated from twitter posts is coherent with the one generated from traditional sources. It was found, that there was an association between rates of suicidal tweets by users and actual suicide rates. No concrete filtering technique was used. But the problem was that Unrelated tweets could also be flagged as suicidal. Another shortcoming being that the study was specific to the US.

Coppersmith, G., et al. talks about the analysis performed on data obtained from OurDataHelps.org and from the Twitter API by using keywords in [8]. The data was classified into different categories - genuine statement, fictitious statement and disingenuous statement. Both supervised and unsupervised models were used to prevent overfitting. The data was projected into a dense vector space. In order to extract contextual information, a bi-directional Long Short Term Memory Model(LSTM) was used. Using skip connections a self-attention layer was developed. For the weights, attention mechanism was used. All of this was fed into a linear layer with a softmax function to predict the result. The results were analysed by plotting ROC curves. This system was able to mitigate overfitting. And using LSTM meant that long term dependencies would be avoided. The analysis was performed on data which predominantly contained data of females aged 18-24 and hence, the system cannot be generalized. The system was optimized in order to detect long term changes in humans as opposed to short term changes which can be unfavorable.

The process proposed in [9] starts with data collected using the Twitter API based on keywords indicating chances of potential PTSD and was segregated into two types - Positive PTSD and Negative PTSD. Three classifiers namely - Unigram Language Model (ULM), One Character N-Gram Language Model (CLM) and one from the LIWC categories were used. ULM performed the best followed by CLM which outperformed LIWC. Collaborating with retired personnel for information about veterans was an added advantage. The LIWC was found to not be able to capture all of the linguistic signals present. Another undesirable part was the data which was obtained from a small demographic region.

The cross-sectional survey conducted in [10] explored the relation between suicidal tweets and history suicidal ideation behaviour. The paper revealed that Young Internet users who have a lifetime history of tweeting "want to die" were 3.2 times more likely to experience suicidal ideation, 3.2 times more likely to have a suicide plan, and 2.1 times more likely to have attempted suicide than those who did not have such any history. Tweeting "want to commit suicide" showed significant relationships with lifetime history of intended self-harm. Thus , several links between Suicidal behaviour and Tweets were discovered.

The paper [11] follows the approach of creation of annotated data , proposing a set of features for classifiers , employing classifiers with proposed features. Annotation of data was done by university students into two broad categories Suicidal Intention Present and Suicidal Intent Absent , each having sub categories.After pre- processing Linear classifiers such as Logistic Regression as well as Ensemble Classifiers including Random Forest (Liaw et al., 2002), Gradient Boosting Decision Tree (Friedman,2002) and XGBoost (Chen and Guestrin, 2016) were employed for classification.Comparison amongst various Performance of classifiers revealed Random Forest to be the most accurate and precise . Including all features like TF-IDF ,SF , POS etc. had the most accuracy. Comparison among classifiers proved to be useful as it gave insights and the features used were exhaustive as all Tweets were covered under the set of features. Still, Pragmatic Difficulty such as sudden topic change and Ambiguous Tweets couldn't be handled by the classifier.

Vioulès, M. Johnson, et al. used a threefold approach in [12]. This consisted of an NLP based text scoring system and a two step ML classifier with C4.5 decision tree (J48) followed by SMO with a Pearson VII Universal Kernel function (PUK). They also used a Martingale Framework on longitudinal data (User based) to calculate change points (detect when the user deviates from typical behaviour). Their approach was very comprehensive and gave a fairly high accuracy, however the datasets were manually annotated and were small in size, thus bringing into question it's reliability.

IV. MACHINE LEARNING FOR SUICIDAL TENDENCY DETECTION

The conventional way of identification of suicidal ideation or behaviour has been done in face-to-face settings or a written test, although these seem ineffective as individuals refuse to communicate freely in such settings, making suicide identification and prevention tough. With the use of social media the problems faced in using conventional methods can be overcome since individuals have started to increasingly share their thoughts and feelings on sites like Twitter, Reddit and other public sharing platforms. Machine Learning is a process which enables automatic identification and

classification using statistical methods to identify patterns and co-relations which exist in data. This process when applied to the Tweets can help us identify people who are suicidal and depressed.The huge amount of data allows a variety of machine learning classifiers to be trained to identify high-risk individuals.

Following are the performance measures used to evaluate the algorithms :

a. **Confusion Matrix**: The confusion matrix is a table used to define the performance of a classifier on data whose true values are known. Terms associated with Confusion Matrix:

i. True Positives: The number of cases where the predicted value was '1' and the actual value was also '1'
ii. False Positives: The number of cases where the predicted value was '1' but the actual value was '0'
iii. False Negatives: The number of cases where the predicted value was '0' but the actual value was '1'
iv. True Negatives: The number of cases where the predicted value was '0' and the actual value was also '0'

b. **Accuracy**: Accuracy of a model is the total number of correct predictions made by it over the total number of predictions made. Accuracy is a good evaluation metric only if the dataset is balanced.

c. **Precision**: It is a measure of how much of the positive classification is actually positive. Ie

$$\text{Precision} = TP/(TP+FP)$$

d. **Recall**: It is a measure of how much proportion of the positive values was correctly identified by the algorithm. Ie

$$\text{Recall} = TP/(TP+FN)$$

e**. F1 Score**: It is the balance between both precision and recall. In other words, it is the harmonic mean of precision and recall.

$$\text{F1 Score} = (2 * P * R)/(P + R)$$

V. DATASET

To implement the algorithms a dataset of 4443 tweets are used. These were collected using Twint, an open source application which lets us access Twitter data without using it's API, thus avoiding its pitfalls. Twint allows filtering of the resulting tweets based on various parameters such as tweet content, username, time, geolocation etc. The tweets gathered were based on tweet content using a dictionary of words which generated based on conversations with psychologists. The tweets are then classified into 7 categories (Table I). The most common words associated with different classes are studied using a word cloud which displays relevant words along with their importance (Fig.1). The result of this word cloud tells us that words such as Depress, Suicide, Want, Feel, Lone are some of the most commonly used and profuse indicators of suicidal ideation. The general sentiment trends of the dataset were also studied to determine which approach would yield the best results (Fig.2). This analysis told us that a vast majority of our data was neutral, followed by negative tweets and the least were positive. In our situation since a non suicidal diagnosis being wrong has a much higher cost than a suicidal diagnosis being wrong, our primary objective was to reduce False Negatives rather than False Positives. This data was then used to train and test the algorithms.The accuracy of the Algorithms is summarized (Fig.3).

TABLE I. DATA CATEGORIES

Class	Name	Description
0	**Suicidal**	A direct intention to kill self
1	**Depressed**	A post that connotes that a person is depressive, since depression is the biggest marker for suicidal ideation
2	**Informative**	A news article or an informative about suicide
3	**Memorial**	A post about friend/family committing suicide
4	**Sarcastic**	Flippant comment about suicide or death
5	**Relevant**	Loosely denote that a person might be sad
6	**Irrelevant**	Contains words related to suicide but aren't actually suicidal

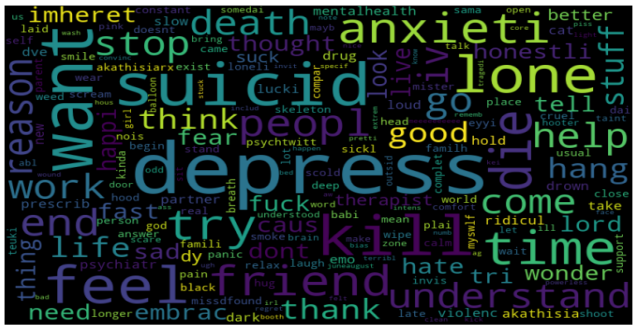

Fig 1: Word cloud of suicidal words from data collected

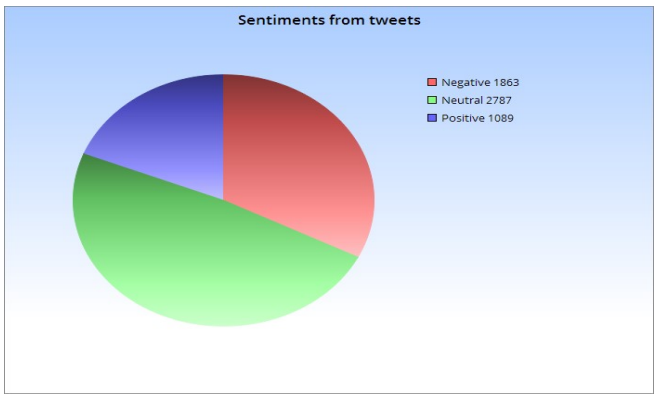

Fig 2: Sentiments from tweets

VI. MACHINE LEARNING ALGORITHMS

The algorithms are studied including Support Vector Machine (SVM), Linear Regression, Logistic Regression, Naive Bayes, LIWC, TF-IDF, Random Forest, GBDT, LSTM, XGBoost, MLFFNN, C4.5 decision tree (J48), SMO. Four of them are chosen which are used often and give consistently high scores:

a. Support Vector Machine : Support Vector Machine (SVM) is a supervised learning algorithm used to distinguish between different classes using data points plotted on a p-dimensional plane . SVM produces a hyperplane which separates the data into different classes , in 2-D this hyperplane would be a straight line . Kernel which is responsible for transformation of problem into linear problem, Regularization which is used to decide the margin of the hyperplane and Gamma which determines the data points to be considered for hyperplane calculation , all these factors are used to tweak the hyperplane in SVM Classifier. When implemented Support Vector Machine on our dataset, got an accuracy of 72 percent.

b. Logistic Regression : This method is popularly used for Binary Classification of data . It is a linear method which uses the logistic function/ sigmoid function to form the predictions.It is also used to produce probabilities of a data point belonging to the two classes used in the model. Maximum-likelihood estimation is one of the most commonly used learning algorithms with this approach . On implementing Logistic Regression on our dataset, got an accuracy of 76.3 percent.

c. Random Forest : This method involves construction of Multiple Decision trees that are based on the 'If- Then' relationship and used for regression and classification .The model chooses Random training data points and Random subset of features during splitting at a node. The main problem with the method is that of overfitting.Since , the algorithm works exhaustively on the training data it captures not only the meaningful or essential relationship between data but also noise. If a maximum depth is assigned, then the error due to bias might increase as the flexibility is reduced.Random forest algorithm constructs a multitude of decision trees and ensembles their results to give the final classification. This algorithm reduces the problem of overfitting as well as the error due to bias. Accuracy of this algorithm increases with the number of trees used. The final predictions are made by averaging the predictions of each individual tree. Implementing Random Forest gave an accuracy of 67.6 percent.

d. Multinomial Naive Bayes : Multinomial Naive Bayes is a specialised version of Naive bayes which takes into account the frequencies of terms and uses conditional probability to classify the text data into different classes. It is mainly designed for classification with discrete features. Multinomial Naive Bayes gave us an accuracy of 72.1 percent.

VII. RESULT

Through the implementation of various ML techniques on the Twitter Dataset it was found that Logistic Regression and SVM perform better than other approaches. Feature selection using TF-IDF yielded improved the effectiveness of the algorithm. The results can be improved by analyzing the data, which words occur most often and what trends do the posts follow. It was found that there were 39 instances of Class 0 signifying Suicidal Tweets and 398 instances of Class 1 i.e Depressed Tweets and 3,393 Tweets were Class 6 , i.e Irrelevant. Clearly the concentration is on a small proportion of the corpus of data. Suicidal Tendency Detection must be further carried out by analysing individuals for putting out these flagged Tweets.

The following table (Table II) gives a comprehensive view of the macro-average results obtained on implementing the aforementioned algorithms on our dataset.

TABLE II. MACRO-AVERAGED RESULTS

Algorithm	Accuracy	Recall	Precision	F1 Score
SVM	**0.720**	**0.182**	**0.204**	**0.183**
Logistic Regression	**0.763**	**0.1687**	**0.297**	**0.170**
Random Forest	**0.676**	**0.185**	**0.202**	**0.171**
Multinomial NB	**0.7212**	**0.122**	**0.111**	**0.113**

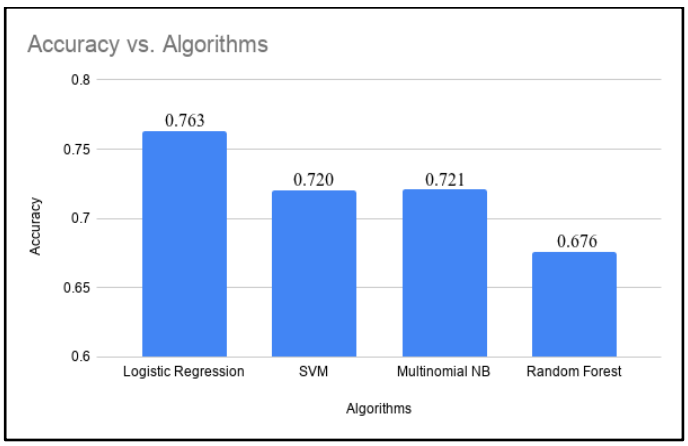

Fig 3: Graph of Accuracies of tested Algorithms

VII. CONCLUSION

Machine Learning applied to detect suicide intention and depression amongst individual is effective as traditional approaches are hindered by factors like face-to-face conversation and shyness to express themselves. In the future like to delve more into context analysis via the retweet history and given links to external sites. Even though ML seems to be

a better method, yet it entails some shortcomings, like eventually having to use human intervention to approve the predictions by the ML model, it's limited to detecting suicidal tendencies and depression and contextual analysis as the data collected doesn't come with prior background explanation.

REFERENCES

[1] "Suicide:one person dies every 40 seconds", World Health Organization 9 September 2019.

[2] Sher, Leo, and René S. Kahn. "Suicide in schizophrenia: an educational overview." Medicina 55.7 (2019): 361.

[3] Ji, S., Yu, C. P., Fung, S. F., Pan, S., & Long, G. (2018). Supervised learning for suicidal ideation detection in online user content. Complexity, 2018.

[4] Vijaykumar, Lakshmi. "Suicide and its prevention: The urgent need in India." Indian journal of psychiatry 49.2 (2007): 81.

[5] Ryu, S., Lee, H., Lee, D. K., & Park, K. (2018). Use of a machine learning algorithm to predict individuals with suicide ideation in the general population. Psychiatry investigation, 15(11), 1030.

[6] O'dea, B., Wan, S., Batterham, P. J., Calear, A. L., Paris, C., & Christensen, H. (2015). Detecting suicidality on Twitter. Internet Interventions, 2(2), 183-188.

[7] Jashinsky, J., Burton, S. H., Hanson, C. L., West, J., Giraud-Carrier, C., Barnes, M. D., & Argyle, T. (2014). Tracking suicide risk factors through Twitter in the US. Crisis.

[8] Coppersmith, G., Leary, R., Crutchley, P., & Fine, A. (2018). Natural language processing of social media as screening for suicide risk. Biomedical informatics insights, 10, 1178222618792860.

[9] Coppersmith, G., Harman, C., & Dredze, M. (2014, May). Measuring post traumatic stress disorder in Twitter. In Eighth international AAAI conference on weblogs and social media.

[10] Sueki, Hajime. "The association of suicide-related Twitter use with suicidal behaviour: a cross-sectional study of young internet users in Japan." Journal of affective disorders 170 (2015): 155-160.

[11] Sawhney, R., Manchanda, P., Singh, R., & Aggarwal, S. (2018, July). A computational approach to feature extraction for identification of suicidal ideation in tweets. In Proceedings of ACL 2018, Student Research Workshop (pp. 91-98).

[12] Vioulès, M. J., Moulahi, B., Azé, J., & Bringay, S. (2018). Detection of suicide-related posts in Twitter data streams. IBM Journal of Research and Development, 62(1), 7-1.

[13] W. Wang, L. Chen, M. Tan, S. Wang, A. P. Sheth, "Discovering fine-grained sentiment in suicide notes", Biomedical informatics insights, vol. 5, pp. 137, 2012.

[14] S. Chattopadhyay, "A Study on Suicidal Risk Analysis," 2007 9th International Conference on e-Health Networking, Application and Services, Taipei, 2007, pp. 74-78.

[15] Sosa, P. M. (2017). Twitter Sentiment Analysis using Combined LSTM-CNN Models. Eprint Arxiv.