International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-2, July 2019

Suicide Prediction on Social Media by implementing Sentiment Analysis along with Machine Learning

K Venkateswara Rao

Abstract: ***Technology is growing day by day and the influence of them on our day-to-day life is reaching new heights in the digitized world. Most of the people are prone to the use of social media and even minute details are getting posted every second. Some even go to the extent of posting even suicide related issues. This paper addresses the issue of suicide and is predicting the suicide issues on social media and their semantic analysis. With the help of Machine Learning techniques and semantic analysis of sentiments the prediction and classification of suicide is done. The model of approach is a four-tier approach, which is very beneficial as it uses the twitter4J data by using weka tool and implementing it on WordNet. The precision and accuracy aspects are verified as the parameters for the performance efficiency of the procedure. We also give a solution for the lack of resources regarding the terminological resources by providing a phase for the generation of records of vocabulary also.***

Key Words: Suicide Prediction, SocialMedia, Machine Learning, Semantic analysis, Classification.

I. INTRODUCTION

Social networks are rapidly changing with the brimming technological updates and are changing the way humans see the world. Everything is on the web and is posted on the social networking websites. These have become a free source of content and is only user driven content and many are getting used to it and believe the posts of a nobody, which can be either true or false. The acquisition of customers on social media is increasing day to day and are like addicted to it as a part of their life. Many are so into it that they forget their surrounding and have been the cause of many incidents and accidents. Of course, these have provided a ground for many to share their views either technical or non-technical but what bothers is that there is no source that can be perfect and true. Each person can only go to an extent to which his point of view can take him/her and we cannot take it as granted and is not always true. Sometimes the user driven content maybe informative and useful but not all the times..

Many suicides can be stopped if predicted and estimated. In this paper, a four-tier procedure to predict the suicides is given in the below sections and is very useful in the prediction and suicides can be minimized if measures are taken to protect the predicted people who can commit suicide. There are situations where some content has led to the deaths of many. Suicide is on the prime issue and as everyone is on the Social networking sites, the flow of information is larger than other modes of communication. Some people even go to the extent of live broadcasting of their suicides and cases where suicide notes are being posted on the social networks

Revised Manuscript Received on 30 July 2019.
***** Correspondence Author
K Venkateswara Rao*, Associate Professor, Department of CSE, VLITS, Vadlamudi, Guntur (dt), (Andhra Pradesh), India.

II. RELATED STUDY

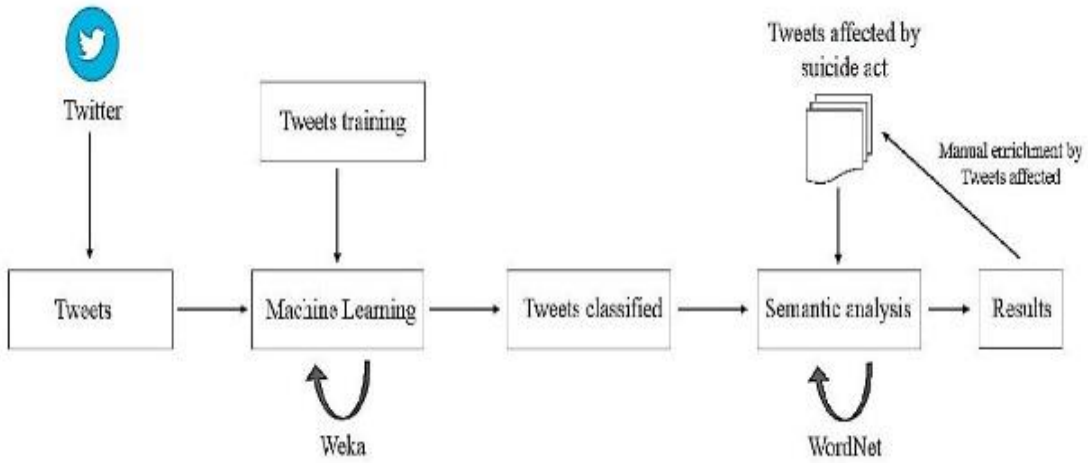
Reference	Variables measured	Comments
Rangel-Garzón et al	Beck Scale for Suicide Ideation (SSI) Beck Scale for Suicide Ideation, Self-Report (SSI-SR) Modified Scale for Suicide Ideation (MSSI) Plutchik suicide risk scale SAD PERSONS scale Suicide Assessment Scale (SUAS) Modified Suicide Assessment Scale (SUAS) and Suicide Assessment Scale, Self-Report (SUAS-S) Suicide Intent Scale (SIS) Adult Suicidal Ideation Questionnaire (ASIQ)	The MSSI and the Plutchik Suicide Risk Scale could be useful in emergency services
Roos et al	Beck Hopelessness Scale Beck Depression Inventory Beck Scale for Suicide Ideation (SSI) Suicide Intent Scale (SIS) SAD PERSONS scale Mini-International Neuropsychiatric Interview (MINI) suicide subscale Suicide Assessment Scale (SUAS) Schedule for Nonadaptive and Adaptive Personality–Self-Harm subscale (SNAP-SH) Karolinska Interpersonal Violence Scale (KIVS) Death/Suicide Implicit Association Task (AIT) Suicide Stroop task	Analysis limited to scales that underwent predictive studies (suicide / attempted suicide) The prediction of future suicidal behavior based on these scales has inconsistent results Neurocognitive assessment tests (Suicide Stroop task, Death / Suicide Implicit Association Task (IAT)) would be more predictive than clinical evaluation
Lotito et al	Beck Depression Inventory Beck Hopelessness Scale Beck Scale for Suicide Ideation (SSI) Motto's Risk Estimator for Suicide Linehan's Reasons for Living Inventory USuicide Severity Scale (C-SSRS) *Chronological Assessment of Suicide – CASE approach* Minnesota Multiphasic Personality Inventory-2 (MMPI-2) Rorschach inkblot test Firestone Assessment of Self-Destructive Thoughts (FAST).	Scales analyzed from the perspective of the American Psychiatric Association (APA) Only included articles in English Does not review the predictive validity of the scales None of the scales predict suicide, but they are a useful tool, along with the clinical interview
Hourani	Beck Scale for Suicide Ideation Reasons for Living Inventory (RFL) SAD PERSONS scale Suicide Risk Assessment (SRA) Potencial Suicide Personality Inventory (PSPI) Suicide questionnaire	Published in 1999 Most scales have poor to moderate predictive capacity. Detects common limitations of the scales: they are based on predictions and not on evaluation models, they cannot be applied to different groups of individuals or settings, they do not incorporate the risk factors in their variables, and they do not consider the effects of interaction between the risk factors. Beck suicide scale prevails as recommended.

III. METHODOLOGY

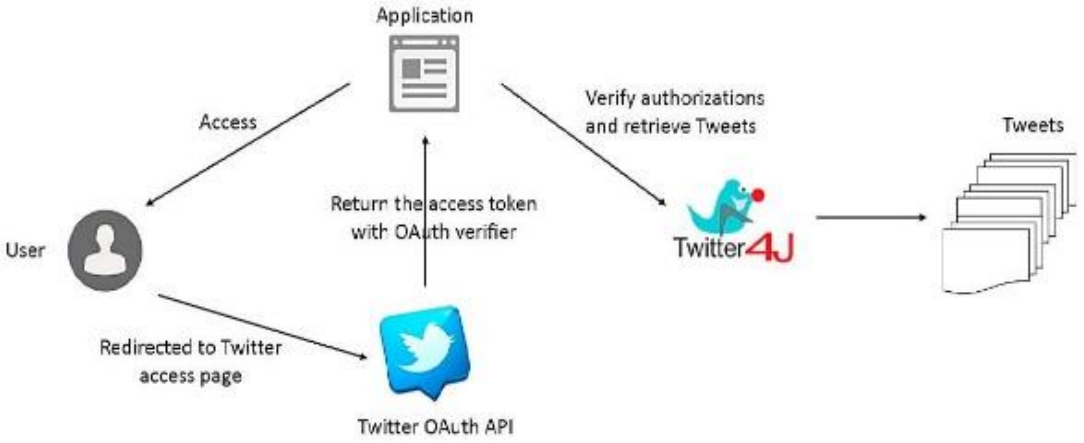
The methodology proposed for the suicide prediction and semantic analysing of the obtained data using WordNet. The procedure is four-tier procedure with systematic order of implementation. The steps of implementation are as follows:

1. Vocabulary diction formation (related to suicide).
2. Data extraction from twitter (Using Twitter4J).
3. Machine learning Technique.
4. Systematic and semantic analysis using WordNet.

The first phase will be of the word dictionary creation, which is related to our theme of study (Suicide). The later phase will be the data extraction from the twitter, which is later used in the procedure of machine learning techniques and semantic analysis. The next step is the machine learning application on the extracted data and predicting the suicide estimation in the users. The last step is the semantic analysis of the obtained result with the help of WordNet. All the four steps together form the four-tier architecture of the procedure for the estimation and prediction of suicides.



Architecture diagram of the procedure of suicide prediction



Data Extraction procedure from twitter data

Machine learning gives us classifiers for data manipulations and other operations. Machine learning also automates the algorithms of classification and its construction. We manually define a training dataset, which consists of the data which has been already infected. The characteristics of the input vectors are distinguished as related classes by the classification model in the Machine Learning. Many classification models are developed in the Machine Learning environment for the simplification of the process with the help of these models.

The WordNet is the diction of English which is represented as the Semantic diction in the procedure as each two words has a relation and the relation is depicted in the WordNet. The WordNet consists of relations of semanticity between words. This acts as the catalyst for the semantic analysis of the acquired relation and depicts the relation between the acquired data and the defined data from the vocabulary created in the first step of the procedure. The semantic measure proposed by Leacock and Chodorow is based on the shortest length between two syntaxes of WordNet. The measure is calculated as

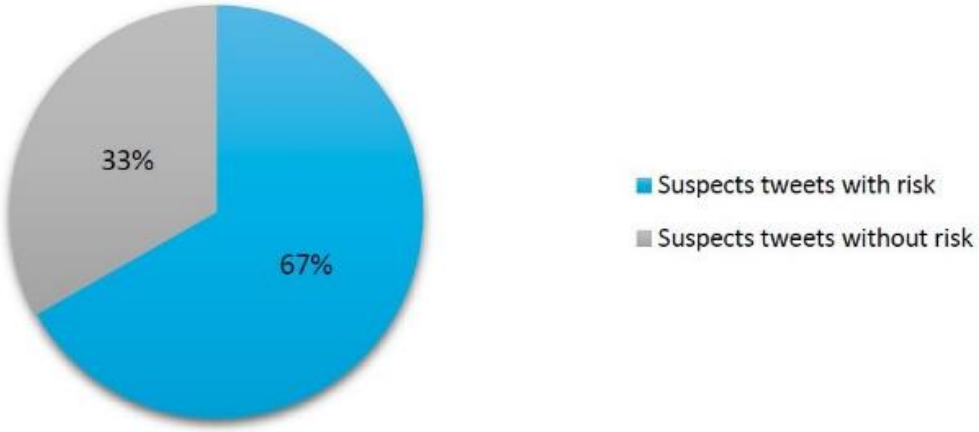
$$Sim_{lc}(a,b) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} a_i * b_j * Sim(i,j)}{\sum_{i=1}^{n}\sum_{j=1}^{m} a_i * b_j}$$

```
InputData : tweetSet, training_tweets
RemoveStopWord (tweetSet) and RemovePunctuation (tweetSet)
For each element ∈ (tweetSet, termOftweet)
Write(tweetSet, termOftweet)
End for
List(IndQuery) = indexing (training_tweets)
SemSim = 0; A = 0; B = 0
For each term ∈ List(termOftweet)
N = wordcount(term)
For each elem ∈ List(IndQuery)
R = wordcount(elem)
A ← A + N × R × Sim(term, elem)
B ← B + N × R
End for
End for
SemSim ← A / B
Return (tweet, SemSim)
```

Algorithm for semantic similarity

IV. RESULTS

The result is obtained when the training data and the result data are in comparison with the help of the above-defined algorithm, which compares the semantic properties of two words. The classification of the data is carried out on the weka tool. Machine learning application (Weka tool) on the extracted data and predicting the suicide estimation in the users. The last step is the semantic analysis of the obtained result with the help of WordNet. All the four steps together form the four-tier architecture of the procedure for the estimation and prediction of suicides. The data extracted is divided such that a part is tfor training set and other is for the test set. The data regarding the result of classifiers is depicted below in the form of a table and the cluster instances which are correct and also incorrect are shown in the bar diagram shown below.



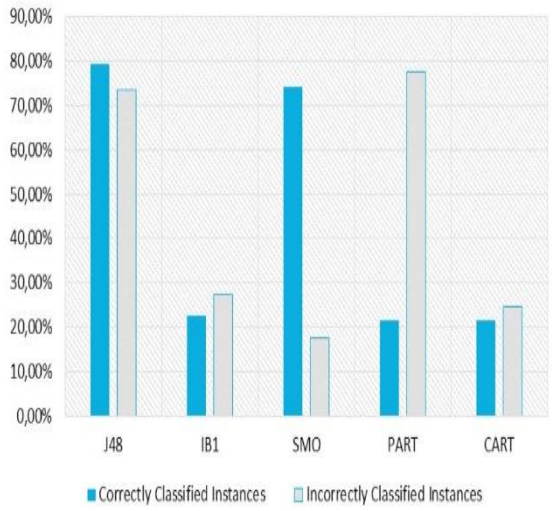
Analysis of tweets with and without risks in the data

Table 1. Cross-validation of evaluations on classifiers for suspected tweets with risk of suicide.

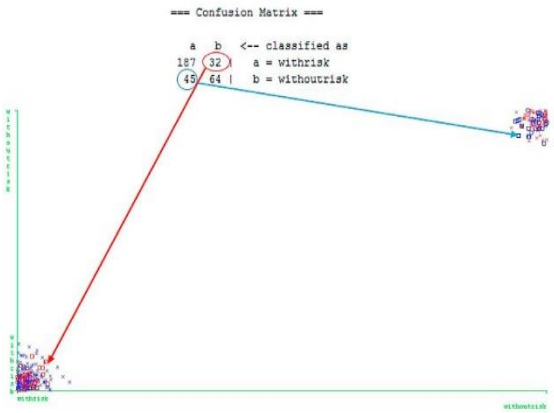
Algorithms	IB1	J48	CART	SMO	Naive Bayes
Precision	77%	81.2%	83.1%	89.5%	87.50%
Recall	82.2%	84.2%	88.5%	89.11%	78.8%
F-score	79.5%	82.6%	85.7%	89.3%	82.9%

Table 2. Cross-validation of evaluations on classifiers for suspected tweets without risk of suicide.

Algorithms	IB1	J48	CART	SMO	Naive Bayes
Precision	63%	75.4%	66.7%	70%	61.00%
Recall	50.5%	72.4%	58.7%	51.4%	76.1%
F-score	55.8%	63.8%	62.4%	59.3%	67.8%



Different classifiers performance stats



Visualization error classifier

V. CONCLUSION

The procedure allows us to implement multiple types of algorithms related to machine learning and by using WordNet we semantically analyzed the data we obtained. This is an initiative against the suicide and further enhancements to the model can be done by improving the efficiency and accuracy factor. Also, the implementation of these on the bigdata framework with multilingual wordnet and machine learning will be the future study for more development and predictions of suicide related issues on the social networking sites.

REFERENCES

1. B. O'Dea, S. Wan, P. J. Batterham, A. L. Calear, C. Paris, and H. Christensen. Detecting suicidality on Twitter. Internet Interventions, vol. 2, pp. 183-188, 2015
2. P. Turney. Thumbs up or thumbs down Semantic orientation applied to unsupervised classification of reviews. Proceedings of the Association for Computational Linguistics.
3. Wilson T., Wiebe J. and Hoffmann P. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In the Advanced Research and Development Activity (ARDA).
4. Kasturi D. V., Nurhafizah T. Suicide detection system based on Twitter. Science and Information Conference 2014, pp. 785-788, August 27-29, London, UK
5. Literature review to identify standardized scales of assessment of suicidal risk in adults seen in primary health care Carolina Abarca, Cecilia Gheza, Constanza Coda, Bernardita Elicer *Medwave* 2018;18(5) e:7246 doi:10.5867/medwave.2018.05. 7246.