International Journal of Psychosocial Rehabilitation, Vol. 24, Issue 05, 2020
ISSN: 1475-7192

Machine Learning- Individual Models verses Ensemble Models on Suicidal Rate

K. Nethravathy[1], D. Saraswathi[2]

Abstract

Sentimental analysis through Machine Learning is a wide area of research in the field of social media. The most popular and universally used social media like twitter helps in gathering the data in all the field of research. The word sentimental points to a very specific feature in the dataset that has to be selected for further analysis. The opinions, thoughts or feelings of the society about one particular topic like suicide/movie/new product can be received by twitter.

The different Machine Learning algorithms like Random Forest, Linear Regression can be considered to check the accuracy. On the other hand the ensemble methods such as Bagging, Boosting and voting can also be applied on the dataset. Hence, individual algorithm and the ensemble methods used are analyzed to estimate most suitable Machine Learning Model The good model gives more accuracy result when applied on realistic life.

Keywords: *Sentimental, opinion, suicide, Machine learning, ensemble, Random Forest, Linear Regression, Bagging*

.

I. Introduction

Sentimental analysis has become a strong tool in many fields of research area especially in data mining. It is mainly detects the opinions of the people on different issues through many social networks like twitter, face book, email etc., Twitter is one of the most widely used and popular micro blogging internet website, through which people post their opinions on the current topics like new products to the society or movie reviews or situation about political developments or stock market news etc., but twitter is really easy, fast and good platform to collect the data for any kind of research. Large set of People directly send their individual opinions, emotions and attitude. The textual form of the message later subjected for preprocessing of data and such opinions are taken into account for sentimental analysis. By using different Machine Learning techniques, it is considered to detect the polarity of text. Later it is categorized as positive, high positive, negative high negative and neutral categories. Machine Learning classifiers like Support Vector Machine, Decision Tree and K-Nearest algorithms are used to classify the data. However, by using the individual classification of Machine learning it is

[1]*Department of Computer Science, Maharani Lakshmi Ammanni College For Women Autonomous (mLAC), Bengalore*
[2]*Department of Computer Science, PSG College of Arts & Science, Coimbatore*

not that effective sentimental classification and accuracy. Hence, an attempt of improving the accuracy and subject polarity it is better to use the multi learning method in sentimental analysis. This leads to hybrid method with a combination of more than one method also known as Ensemble Learning. This would give a higher developmental accuracy classification.

The twitter messages are text formats opinions taken into "bag of words" (BOW) for synonyms and synset of word that exist. Such words can be later differentiated into parts of speech (POS). [effect of Negation] The ensemble learning includes many methods like Voting, Boosting, Bagging, Staking, cascading etc., The major types that is suitable are Voting, Bagging and Boosting for text based sentimental data analysis. This is and extended development of Natural Language Processing (NLP). The Bagging type of ensemble is popularly used text classification. Since sentimental analysis in suicidal cases gives different variety of thoughts, emotions, it becomes very ambiguous to predict as positive, negative or neutral for a particular word. However, the breaches of sentimental extraction can be ignored and only valid words are to be collected. In addition to this it is important to consider the evaluation of suicide sentimental analysis to find the comments and suggestions about suicide. So, which Sentimental Analysis methodis best suitable for the extraction of subjective polarity in suicide, we need to survey?

In this work it is to make a proposal about suicide data from southern states of India. There are different reasons for suicide and each one has its own validity but the more effective reason on the society has to be focused for analysis. After the accuracy with Machine Learning models we can get the prediction in the coming future. so such situation can be handled in advance to avoid or to bring down the rate of suicide.

The article is organized into Five Sections. Section 1 has Introduction – about the work carried out. Section 2 includes Literature Survey relating to title of the article. Section 3 –Discussions and Classification algorithms Section 4. Experimental /Comparative Results of suicide sentiment analysis and the classified sentiments generated. Last Section 5 Conclusion and Future scope of the work.

II. Literature Survey:

Different Machine Learning algorithms were discussed on different datasets by [1] Twitter datasets were divided into three major types such as positive, negative and neutral about new product reviews by the customers. The twitter messages are also divided into three levels such as word level, phrase level and document level. Earthquake data set were considered and to detect the real time changes. They proposed a model to monitor and detect the changes like earthquake by considering the twitter user as sensor [1.1] Other Sports related twitter messages were taken into account to find out the hidden time most important actions in the event. Hidden Markov Model was used to find out the hidden events actions. Other three other algorithms were also used to summarize the tweet messages. [1.2] Sentimentanalysis on unbalanced datasets are like languages of Arabic and English. SVM, NB algorithms were applied and g-performance technique was adopted for evaluation measure. [1.3] sentiment analysis on text to speech based on the features like positive, negative and neutral. Through corpus the data was trained and tested and classified as unigram to get more accurate result. Finally, all the various Machine Learning methods for sentiment analysis on different types of data sets like real time, event detection, unbalanced were summarized. It is stated that the SVM and Naïve Bayes techniques were intensively suited to get better accurate results. How supervised learning of ML technique on sentiment analysis

of twitter messages works by applying the Apriori algorithm fails was discussed [2] It finishes in generating the erroneous results. This was overcome by using Support Vector Machine algorithm in statistical result on a particular field. The irrelevant feature selection was eliminated and high dimension feature with some properties were considered. and they were generalized. It is concluded with SVM over ANN algorithm. Chakrit Pong-Inwonget.al [3] discussed on teaching evaluation by improved sentiment analysis by using the ensemble learning algorithm. The feedback of data was taken for data preprocessing with sentiment feature selection. The Voting Ensemble algorithm was experimented to reduce the attribute. The higher prediction factor was resulted in arriving good accuracy. Negation effects in sentiment analysis [4] certain words relating to negation were identified and calculated to improve the classification. The sentences with and without negation were considered for analysis. The accurate polarity of the negation words smoothen the work in improving the precision. Hence the experiment showed in producing a significant improvement in result like recall and accuracy.

ML algorithms like Hoeffding tree and McDiarmid tree were discussed on twitter messages [5] in sentiment analysis. The emotions, opinions, detection etc., were intend to recognize the polarity of the twitter messages. Filtering and wrapper preprocessing techniques used to enhance the performance of features and proper classifying algorithm was applied Both the mentioned algorithms result in very close accuracy but the time consumed for analysis was high in Hoeffding tree but less in McDiarmid tree. Arabic Sentiment Lexicon was discussed in building an Automatic Lexicon Expansion method [6]. The expansion was experimented as first iteration polarity distribution and second iteration polarity distribution. The lexicon techniques generated removed the repeated words, special characters and English words. In the aggregated lexicons the polarity was less with that of expanded lexicon which leads to improved accuracy level. The work concludes that the un-stemmed words leads to proper Machine learning sentiment analysis of lexicons. This gives good performance of accuracy and prediction but stemmed words with less polarity decreases the accuracy. Opinion Mining in Big data [7] can be of supervised and unsupervised data. To apply the Machine Learning methods like Naïve bayes, Support Vector Machine on the data set selected may not be suitable for opinion detection.But sometimes a single algorithm may not be able to reach the expected result. The Hybrid approaches like BOW, POS, SVM etc. can be combined so as to get better opinion extraction and sentiment analysis of big data. Ensemble Classifier is an effective method of ML [8] for sentiment analysis. Comparative table showing the different algorithms used on the UCI dataset results in accurate percentage. The NB, SVM when used singly shown less accuracy and the hybrid/ensemble algorithm like voting, bagging bootstrap methods resulted in high improvement of accuracy.

III. Discussion and Classification algorithms

In the architectural diagram it is shown from data collection till the Result ie., final model with accuracy. It is easy to understand the flow of work and the in between steps to complete the sentiment analysis approach. It involves few steps and inn each step the data gets processed and analyzed to attain a significant prediction.

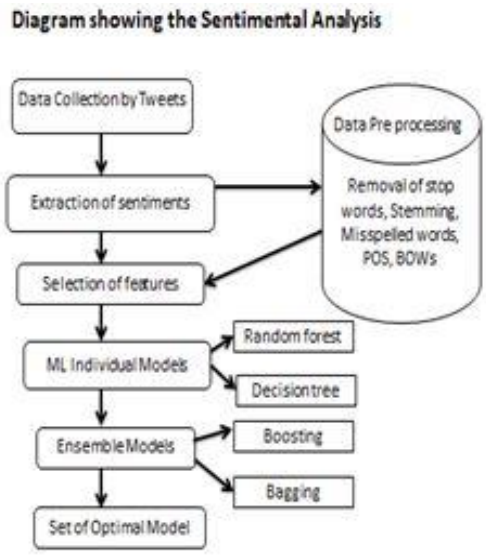


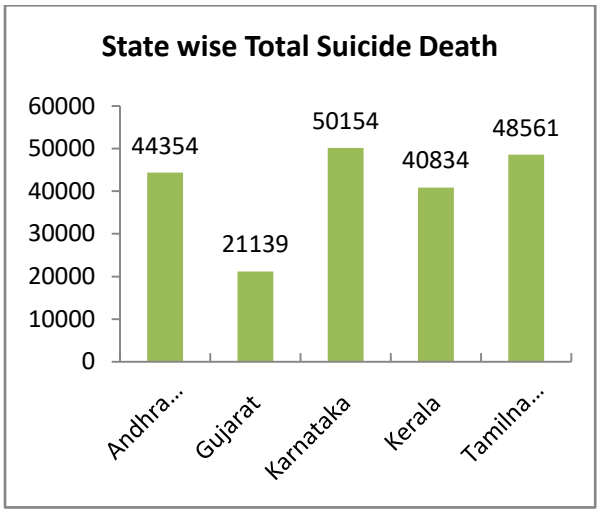
Fig.1 Diagram showing Seentimental Analysis

Data collection: The data is collected from the www.kaggle.com web site in the form of .csv file. It is later converted to excel for further analysis which is suicide data for five years from 2007 to 2012. The size was too huge and hence decided to focus only on south state suicidal rates. Five states were identified with high suicide reasons. The top reasons like Love affair, Agriculture Activity, Education Status, By Hanging and Never Married. It is filtered with total death in each state as shown in Table1.

States	**Total Death**
Andhra Pradesh	44354
Gujarat	21139
Karnataka	50154
Kerala	40834
Tamilnadu	48561

Table.1

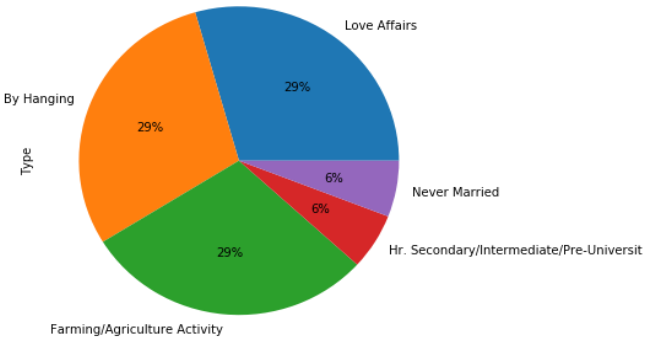
The graph is drawn showing the clear comparison of highest death among five states. Karnataka tops with the number 50154. (graph1).



Graph.1- State wise total suicides

Data preprocessing: Five year from 2008 to 2012 of five south states namely Andhra Pradesh, Gujarat, Karnataka, Kerala and Tamilnadu is filtered from main CSV file. It has a total of 851 data sets which is further considered for data preprocessing by using python coding. It is done using Corpus and feature extraction is finalized. The selected feature is subjected for processing. The Processing involves removal of multiple spaces, single character, conversion of lower cases etc. After preprocessing the data is trained and tested for classification.

A Pie chart was plotted from the entire data of csv file (pie chart1). It shows clearly form which state more suicide is recorded and the appropriate reason for suicide.



Pie Chart.1- Succide Data with reasons

Total of male female of all types of reason for suicide is represented in table2.

```
southstates.Gender.value_counts()
```

```
Female    425
Male      425
Name: Gender, dtype: int64
```

Table.2

It has a total of 850 data of both the genders which includes all types of reasons for suicide. It can be decided that both the gender have the same ratio of death rate. The Random Forest classification algorithm is

used the analysis of suicide data. It gives the following report, showing the Classification Report, Confusion matrix and Accuracy. The Classification Report has Precision, Recall, F1 square and

*Data visualization:*by using SVC classifier along with built in libraries like pandas, seaborn, metplotlib tools. In this paper it is yet to make another specific feature along with suicide death rate.

IV. Experimental/Comparative Results

```
from sklearn.metrics import classification_report, conf

print(confusion_matrix(y_test,predictions))
print(classification_report(y_test,predictions))
print(accuracy_score(y_test, predictions))
```

```
[[51 31]
 [64 24]]
              precision    recall  f1-score   support

      Female       0.44      0.62      0.52        82
        Male       0.44      0.27      0.34        88

    accuracy                           0.44       170
   macro avg       0.44      0.45      0.43       170
weighted avg       0.44      0.44      0.42       170

0.4411764705882353
```

Fig.2

The accuracy mentioned in Fig.2 is 0.44 from Random Forest as individual algorithm. Other algorithms can also be checked for prediction and accuracy for more improvement of performance. but the ensemble algorithms are more effective in improving the results. The Gradient Boosting classifier and Ada Boost classifier algorithm gives more appropriate accuracy (Fig.3 and Fig.4) respectively.

```
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.model_selection import train_test_split
X, y = make_classification(random_state=0)
X_train, X_test, y_train, y_test = train_test_split(X, y, random_st
clf = GradientBoostingClassifier(random_state=0)

clf.fit(X_train, y_train)
GradientBoostingClassifier(random_state=0)
clf.predict(X_test[:2])
clf.score(X_test, y_test)
```

```
0.92
```

Fig.3 Gradient Boosting Classifier

```
from sklearn.ensemble import AdaBoostClassifier
X, y = make_classification(n_samples=1000, n_featu
clf = AdaBoostClassifier(n_estimators=100, random_
clf.fit(X, y)
AdaBoostClassifier(n_estimators=100, random_state=
clf.predict([[0, 1, 0, 1]])
clf.score(X, y)
```

```
0.983
```

Fig.4AdaBoost Classifier

V. Conclusion and Future works

It is experimented with Machine Learning algorithms on suicide data for sentiment analysis. The Reasons for suicide were considered as the main feature after data processing. Data train for the relevant classification and prediction through accuracy. Comparative accuracy generated by different algorithms is given in Table3.

Random Forest	Gradient Boost	Ada Boost
0.44	0.92	0.98

Table.3 Accuracy comparison

Random Forest classifier gives the accuracy performance is least compared with ensemble algorithms like Gradient Boost and ADA Boost. Hence the hybrid algorithms or ensemble algorithms are more effective rather than an individual algorithm. There are chances of reaching an erroneous conclusion by using single algorithm in Machine Learning.

Future works: More refining of the data set can be done like gender for each specific suicide reason in single state. This can be further visualized by using different techniques like boxplot or facet grid methods for the better understand of the data and more accurate analysis.

REFERENCES

[1] BholaneSavitaDattu “Asurvey on Sentiment Analysis on Twitter dta using different techniques” IJCSIT Vol6(6) ISSN0975-9646 201

[2] AnkitPradeep Patel et..al “Literature Survey on Sentiment Analysis of twitter dta using ML approaches” IJIRST Vol(3) 2017 ISSN2349-6010

[3] Chalrit Pong-Inwong “Improved Sentiment analysis for teaching Evaluation using feature selection and voting ensemble learning integration” IEEE 2016 NO.978-1-4673-9026-2/16

[4] Wareesa Sharif et..al “Effect of Negation in Sentiment Analysis” INTECH -2016 IEEE No.978-1-5090-2000-3/16

[5] Zahra Rezaei, MehrdedJalali “Sentiment analysis on twitter using McDiarmid Tree Algorithm IEEE 2017 No.978-1-5386-0804-3/17

[6] Mohab Youssef, Samhaa R. El-Beltagy “MoArLex: An Arabic Sentiment Lexicon Built through automatic Lexicon Expansion” Pocedia computer science 142(2018) 94-103

[7] Uma Gurav, Dr. Nandinisidnal “Opinion mining for reputation evaluation on unstructed Big Data” IJARCET Vol (4) 2015 ISSN:2278-1323

[8] Isha Gandhi, MrinalPandey “Hybrid Ensemble of Classifiers using voting” 2015 IEEE No978-1-4673-7910-6/15