ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 5, Issue 2, February 2018

Reliable Prediction for Detection of Heart Attack among People Using Twitter

[1] V. Diviya Prabha, [2] R. Rathipriya

[1] Research Scholar, Department of Computer Science, Periyar University Salem-11

[2] Asst.Professor, Department of Computer Science, Periyar University Salem-11

***Abstract* - Social Media is a powerful tool to gather public personal information through online. It acts as a mediator between patients and general people to communicate and share thoughts. The survey results that in South India nearly around 1.6 million are suffering from heart disease which leads to death. Data is taken from the twitter and prediction for detection of heart attack is obtained by logistic regression to create general awareness in public.**

Keywords: Prediction, twitter, regression, awareness, social media, heart attack, diabetics, blood pressure.

1. INTRODUCTION

In India, the most common disease is heart attack which leads to death. It is a disease that we cannot identify in advance in medical fields. It is the familiar task that people share their thoughts and ideas in public like Facebook, Twitter, etc., Among which Twitter plays an important role helps to access the input data available in the Twitter database. It is important to mining the data in the web in today's world. The health-related data are very sensitive information that also contains personal details. The Web has become a mediator to communicate and suggest certain ideas for the people. There are many reasons to extract data from Twitter

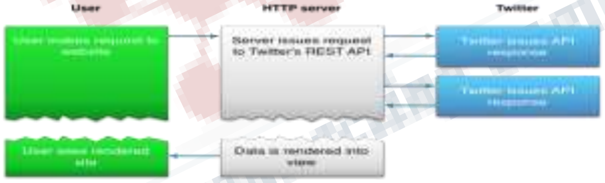

User
HTTP server
Twitter
User makes request to website
Server issues request to Twitter's REST API
Twitter issues API response
Twitter issues API response
User sees rendered site
Data is rendered into view

Fig 1: Streaming of API in Twitter

It nearly performs 140 health related data uses to people. Data collection is an important role for the users [1] Twitter performs the task easier. It has streaming API (Application Program Interface) is a web-based software tool. It makes the user and Twitter interaction to access only needed information by the users. It helps to read and write data on Twitter. The below diagram is represented by Twitter on streaming API.

First, a user can sign in to twitter to make the request for streaming API. REST API is used to read and write Twitter data it will identify Twitter application responses will be available in JSON. Install tweepy in python [2] to communicate with the Twitter platform. Then we need permission to access tokens from http:// www. dev. twitter. com/ .The below figure delivers a message to access tokens as a health-related records app. The streaming API consist of three types of streams they are [17]: User stream –tweets of user, Public stream-tweets of public and site stream-tweets from multiple user. This health care data site stream is used to get data from different users from various domains. The following figure shows how the twitter allows individuals to access data from the database using the secret key.

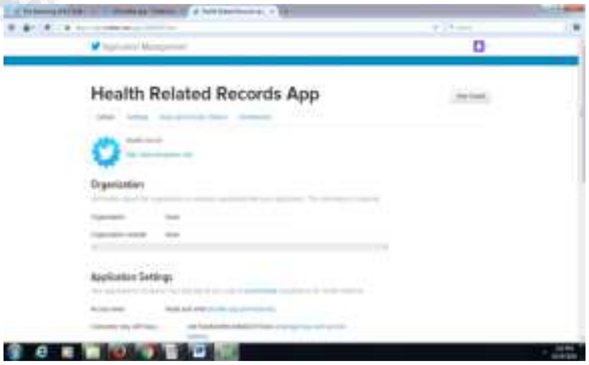


Fig 2: API key in Twitter

It represents a unique consumer key (API key), and it also gives an access token, accesses token secret key to retrieve the Twitter data. From which the twitter data are red and write data these are steps to access Twitter data using API.

Prediction using logistic regression [4] is a kind of linear regression to make a relationship between dependent variable and independent variable. Nearly huge amount of people die due to cause of heart attack every year. Early prediction will help to reduce the death and increase the life history of people. This paper consist of two methodologies: first is to predict the heart disease among the people uses twitter and second is to obtain the accuracy level of the prediction

3.OBJECTIVE

The main objective of this paper is to create awareness among the people who have high diabetics, and blood pressure will have a chance of getting the heart attack. So that people may improve the medicine and treatment for future concerning with the doctor. This is an efficient technique to know the diagnosis before it takes place. It helps to take necessary remedial measures. The accuracy is improved using proposed to binarize logistic regression technique.

4. RELATED WORK

Logistic regression [3] is one of the common data mining techniques used to predict the event of the occurrence or not. Here whether or not the person with symptoms like diabetics, blood pressure, sick, etc., have a chance of getting the heart attack. It takes the data in a binary format of 0 and 1.The Genetic algorithm [14] with logistic regression is a key for prevention or slowing of disease Alzheimer's disease. Data mining [15] using classification algorithm heart disease of patients is predicted. Predictive performance of estimates and different classifier algorithm is carried out.

Classification tree algorithm is used like [16] J48, Logistic Model Tree Algorithm and Radom Forest for which the results are computed represents the method of managing categorical variable in machine learning.

5. METHODOLOGY

In this paper health care, dataset is taken from the question answered by the users online on Twitter. Princeton Survey International has given the data collected from the Twitter. The data present here is the categorical value of type question and answer yes/no to deal these values logistic regression is used. Prediction of the dichotomous result is made. Patients having blood pressure and diabetics predict to cause heart prediction of various other diseases also.

2.2 LOGISTIC REGRESSION

In a logistic regression, a categorical [4] variable is predicted from variables one or more independent in this Twitter dataset cause of heart attack is predicted from blood pressure and diabetics' independent variable. It is used to develop a predictive model of binary in nature. A binary regression is by 1 (symptom to cause a heart attack) and 0 (no cause of heart attack).

$$\pi_i = \Pr(Y_i\,|X_i = x_i) = \frac{\exp(\beta_0+\beta_1 x_i)}{1+\exp(\beta_0+\beta_1 x_i)} \qquad (1)$$

$$logit(\pi_i) = \log(\frac{\pi_i}{1-\pi_i}) \qquad (2)$$

The equation 1 and 2 represents logistic regression variable π_i be the single categorical variable for i users. β_0 and β_1 be the patients having blood pressure and diabetic disease. Using the logit function the heart diseases value is predicted.

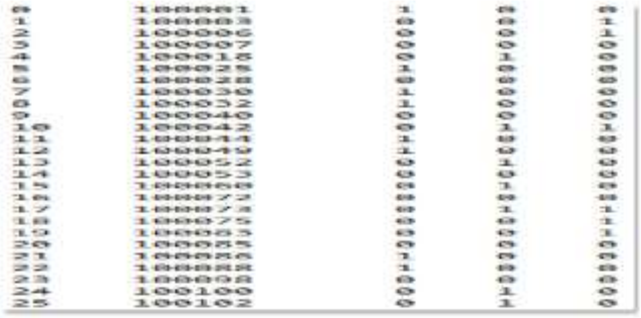
2.3 BINARIZE LOGISTIC REGRESSION

It is the proposed method the process data cleaning and binarize method by fixing threshold value to 0.5 so that it seems to improve the accuracy level of data for given dataset.

$$\pi_i = \Pr(Y_i\,|X_i = x_i) = \frac{\exp(\beta_0+\beta_1 x_i)}{1+\exp(\beta_0+\beta_1 x_i)}(\,threshold) \qquad (3)$$

```
 if    $\beta_0, \beta_1$ = 1 then
      threshold = 1
else
       threshold =  0
end
```

The equation 3 depicts same as that of logistic regression except that it has a threshold value to be placed to improve the prediction techniques and accuracy calculation. β_0 and β_1for blood pressure and diabetic if their value is equal to 1 depicts the threshold value to 1 else to 0.



```
0    100001    1    0    0
1    100003    0    0    1
2    100006    0    0    1
3    100007    0    0    0
4    100018    0    1    0
5    100025    1    0    0
6    100028    0    0    0
7    100030    1    0    0
8    100032    1    0    0
9    100040    0    0    0
10   100042    0    1    1
11   100044    1    0    0
12   100049    1    0    0
13   100052    0    1    0
14   100053    0    0    0
15   100068    0    1    0
16   100072    0    0    0
17   100073    0    1    1
18   100075    0    0    1
19   100083    0    0    1
20   100085    0    0    0
21   100086    1    0    0
22   100088    1    0    0
23   100098    0    0    0
24   100100    0    1    0
25   100102    0    1    0
```

Fig 3: Binarize Logistic Regression

Table 1: Comparison of predicted heart attack people on Twitter

No. of Users	Male	Female	Diabetics Patients	Blood Pressure	Predicted using Logistic Regression	Predicted using Binarize Logistic Regression
N=3014	1337	1677	374	895	260	754

The above figure 3 describes the snapshot of binarize logistic regression of which it is turned as binary values 0 and 1.If a person has blood pressure it is taken as 1 and 0 otherwise similarly for diabetics. Table 1 suggest the number of the individual in twitter is taken N=3014 for these users 374 have diabetic, and 895 have blood pressure out of which it is predicted that 754 people have the chance of getting the heart attack, so the treatment and discussion with the doctor will help to avoid heart disease. Whereas our logistic regression predicts 260 a person which is less compared to binarize logistic regression.

2.4 ACCURACY CALCULATION

It can be performed in two steps. The first step is preprocessing the data the values from the health care data set is converted to binary value to perform the efficient binary logistic regression. Formula [5] for logistic regression let Y is a binary variable.Table 2 describes the accuracy of predicted data calculation of precision, recall and f1-score and support. Formula for [13] for calculating accuracy

$$Presicion = \frac{TP}{TP+FP} \qquad (4)$$

$$Recall = \frac{TP}{TP+FN} \qquad (5)$$

$$f1 - Score = 2.\frac{PRECISION.RECALL}{PRECISION+RECALL} \qquad (6)$$

The equation TP represents True Positive for patients having the prediction of a heart attack. Similarly, FN represents False Negative in equation 5. The equation 6 performs the fi-score accuracy suggestion. The below accuracy table describes the data predicted is, and the value is true to the satisfaction.

Table 2: Accuracy calculation for predicted data

Classes	**Precision**	**Recall**	**F1-Score**	**Support**
0.0	0.77	0.87	0.82	162
1.0	0.71	0.55	0.62	92
Avg/Total	0.75	0.76	0.75	254

Table 2 depicts the output of logistic regression accuracy for two classifiers that 0 represents for low –risk patients and 1 represent for high-risk patients in the prediction of the heart attack. Among the 3014 people, it predicts 254 have the chance of getting the heart attack.

Table 3: Accuracy calculation for binarize logistic regression

Classes	**Precision**	**Recall**	**F1-Score**	**Support**
0.0	0.90	0.95	0.92	630
1.0	0.63	0.45	0.53	124
avg/total	0.85	0.87	0.86	754

Table 3 describes binarize logistic regression when compared to the logistic regression it accuracy level is increased to 0.86 percentage. Among the 3014 people, it predicts 754 people having the chance of getting the heart attack. With it 630 patients have low-risk, and 124 predicts high-risk of the heart attack.

2.3 WORKFLOW

The below figure 4 shows the flow of proposed work for prediction of heart attack among the people on Twitter.

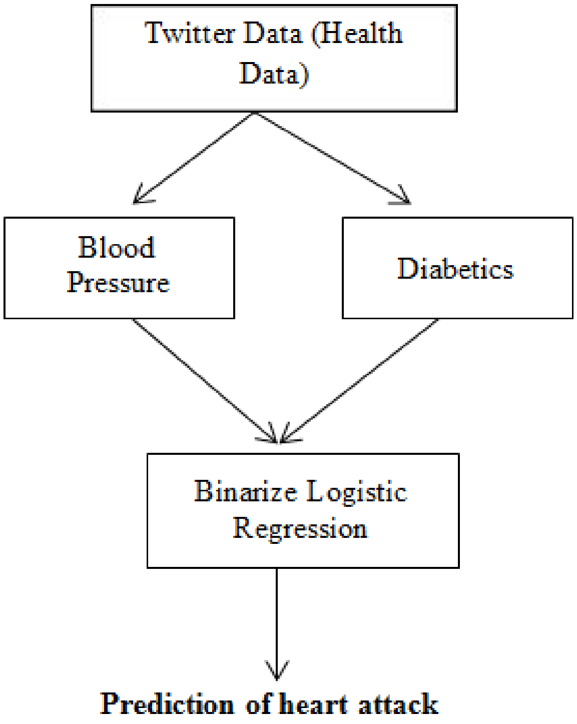

Twitter Data (Health Data)
Blood Pressure
Diabetics
Binarize Logistic Regression
Prediction of heart attack

Fig 4: Flow of proposed work

The Twitter data is extracted from the Twitter website then data is preprocessed with the identification of blood pressure and diabetic with these symptoms it predicts the heart disease. With this, binarize logistic regression techniques for the variable diabetics and blood pressure of a people heart disease is predicted.

3. CONCLUSION

The above predicted result shows that the detection of heart attack for users among 3014 users 754 can detect a heart attack is a suggested output from the logistic regression function. It creates awareness and care to Twitter users and takes necessary treatments. This is an initial step of research using prediction provides an accurate result. The dataset details after preprocessing the healthcare data.

4. FUTURE WORK

The future work is in the variable selection of large dataset is quite difficult with this method. So it is possible to use quick reduct algorithm or some feature selection technology. With this, it reduces the time and produces the result as efficient as possible. Mining in Twitter is another task to get the real data and perform the work easier to predict more and create awareness to reduce the cause of disease.

REFERENCE:

[1] Haewoon Kwak, Changhyun Lee, Hosung Park and Sue Moon (2010). What is Twitter, a Social Network or a News Media? ¸We make our dataset publicly available online at http://an.kaist.ac.kr/traces/WWW2010.html

[2] Abhishanga, Upadhyay Luis Mao (2014). Mining data from Twitter

[3] T.J. Cleophas and A.H. Zwinderman, Machine Learning in Medicine(2013), DOI 10.1007/978-94-007-5824-7_2, © Springer Science Business Media Dordrecht

[4] Sreejesh, Sanjay (2013). Chapter 11,"Binary Logistic Regressions". Sreejesh et al., Business Research Methods,DOI:10.1007/978-3-319-00539-3_11,@Springer International Publishing Switzerland pp :245-258.

[5] G. Rodiguez (2007), Chapter 3,"Logit Models for Binary Data". Pp 3-50

[6] Archer, K. J., S. Lemeshow, and Hosmer, D. W., (2007). The goodness of fit tests for logistic regression models when data are collected using a complex sampling design. Computational Statistics & Data Analysis 51

[7] T. Zaman, R. Herbrich, J. Van Gael, and D. Stern (2010) Predicting information spreading on Twitter. In NIPS Workshop on Computational Social Science and the Wisdom of Crowds,

[8] Mir Sajjad Hussain Talpur, (2013). The Appliance Pervasive of Internet of Things in Healthcare Systems" School of Information Science & Engineering, Central South University (CSU), 410083 - Changsha, China.Vol 10 Issue 1, No 1.

[9] Walter FM, Emery J, Braithwaite D, Marteau TM. (2004). Lay understanding of familial risk of common chronic diseases: a systematic review and

synthesis of qualitative research. Ann Fam Med;2(6): 583–94.

[10] Songhua Xu1*, Deng, PhD; Christopher Markson (2016) Leveraging Social Media to Promote Public Health Knowledge: Example of Cancer Awareness via Twitter, S, JMIR Public Health And Surveillance.

[11] Li Xiang-wei, Qi Yian-fang, (2012), A Data Preprocessing Algorithm for Classification Model Based On Rough Sets, Published by Elsevier B.V. Selection, doi: 10.1016.

[12] Alireza Baratloo, Mostafa Hosseini, PMC (2015).Part 1: Simple Definition and Calculation of Accuracy, Sensitivity and Specificity “ [Online]: https://en.wikipedia.org/wiki/Precision_and_recall

[14] Piers Johnson, Luke Vandewater, BMC “Genetic algorithm with logistic regression for prediction of progression to Alzheimer’s disease”, Johnson et al. BMC Bioinformatics 2014, 15

[15] Hlaudi Daniel Masethe, Mosima Anna Masethe (2014),”Prediction of Heart Disease using Classification Algorithms”, Proceedings of the World Congress on Engineering and Computer Science Vol II, 22-24

[16] Jaymin Patel, Tejal Upadhyay ,(2015), Heart Disease Prediction Using Machine Learning and Data Mining Technique, Volume 7.Number 1,pp129-137.

[17] Shamanth Kumar Fred Morstatter ,(2013), “Twitter Data Anaytics”, Springer.

[18] Godswill Chukwugozie Nsofor , (2006), “Comparative Analysis of Predictive Data-Mining Techniques “,