Analysis of Diabetic Awareness among the General Public in Tamilnadu using Social Networking Data

J. Ramsingh
PhD Research Scholar
Department of Computer Applications
Bharathiar University
Coimbatore, Tamil Nadu, India
j.ramsingh@hotmail.com

V. Bhuvaneswari
Assistant Professor
Department of Computer Applications
Bharathiar University
Coimbatore, Tamil Nadu, India
bhuvanesh_v@yahoo.com

***Abstract*- Big data has rapidly developed into a hot topic that attracts extensive attention from various domains such as industry, government, health care, agriculture and in many sectors. Big Data inexorably draws attention of many data analysts from various countries around the world, in India there exists a huge healthcare burden in terms of diabetes due to rapid urbanization, life style changes. India faces several challenges in diabetic management due to the lack of disease awareness, raising prevalence of diabetic complication among the public. Many barriers prevail among the patients, due to the health care systems that exist; analysis of diabetic awareness must be carried out to make India healthier. New technologies have evolved which serves as a tool to engage and involve patients in health care. This paper gives about the role of big data analytics and Hadoop, in revealing the awareness various aspects of Diabetes Mellitus (DM) is essential for the prevention, management and control of the disease. However, several studies have consistently shown that awareness of DM in the general population is low. This condition constitutes a major public health problem in the country. By the means of data collected through the social network WhatsApp.**

Keywords- Hadoop, Big Data Analytics, International Diabetes Federation, terabytes, geophytes, WhatsApp, Social networking, Diabetic

I. INTRODUCTION

Today we are in digitized world, with this digitalization the current trend in healthcare industry leads to the generation of huge volume and variety of data (Big Data). Big Data can be easily referred to as data which is huge, but more importantly Big Data comes from multiple sources rather than just one, the generation of Big Data leads an analytics called Big Data Analytics [1]. Big Data analytics is advanced analytic techniques used to analyze different types (structured, unstructured and semi- structured) and size of data (terabytes to geophytes) [2], [3]. Analyzing big data allows analysts, researchers, and business users to make better and faster decisions using Big Data that was previously inaccessible or unusable. Were came a slogan Big Data, Big money.

Big Data Analytics in healthcare contribute a major role in processing and analyzing the data in variety of forms to deliver suitable insights [4], [5]. The increased use of Social networking among public helps doctors to reach out to patients, guide them for treatments, provide counseling, creates close-knit support communities and faster recovery. Numerous blogs has been created and may users share large quantity of vision about a particular topic. When Diabetes Mellitus is taken in to account many blogs, social networking sites, mobile apps contribute to the generation of big data. Diabetic Mellitus the most common metabolic disorder in the world, according to International Diabetes Federation (IDF) the number of people affected with diabetes mellitus will increase to 552 million by 2030, over twice the number in 2000, From the survey it is estimated that nearly 21% of these new cases will be from India, which has the highest number of cases in any country [6].

Anne Cooper and ParthaKar [7] stated that people with diabetes build up expertise in self-management through day-to-day living with the resource available for people with diabetes through social networks, blogs and patient self-help. Stafford et al stated that people with diabetes spending a varying amount of time on self-care, with an average of about 20 minutes per day. In India currently 61.3 million people where affected with diabetics, a survey says that peoples in India affected with diabetes will increase to 103 million by 2030 and India will be called the "Diabetes Capital" of the world .The Diabetic disorder occurs due to the life style changes, lack of physical exercise, Imbalanced diet and Genetic disorders.

The awareness of diabetic among Indian's are very poor and many are not even diagnosed with diabetics [8]. There exist many diabetic myths among Indians like "Occurs due to lack of physical labor, Possible only after 50 year of age, Diabetic II is considered as rich man diseases". Several studies from different regions of India have shown that the Type2 Diabetic Mellitus is increasing – from 8.2% in 1992 to 18.6% in 2008 for urban areas, and from 2.4% in 1992 to 9.2% in 2008 in rural areas. A study of epidemic is necessary to understand the risk factors of Diabetes Mellitus. The objective of the work is to analysis awareness on diabetes using Big Data technology. The dataset were collected using social media (WhatsApp). This analytics makes use of 1300 instance of data, which comes under Variety concept in Big Data. The analytics is helpful to analysis the awareness about diabetes and its risk factors using R-Hadoop.

II. FRAMEWORK AND METHODOLOGY

To accomplish the objective a framework is designed with three phases the figure 1 represents the overall framework.

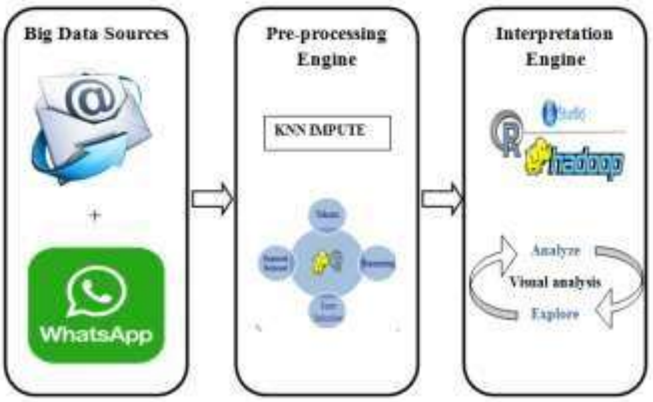


Figure 1. Diabetes awareness analyzer

The first phase is the extraction of data from the social networking sites, the second phase is to preprocess the unstructured data into structured format, the third phase is to interrupt the results from the data collected using exploratory data analysis methods.

A. Data Source

Health care data is of different form which is classified as internal and external sources. The data generated is of multiple formats and resides at multiple locations; sources of healthcare data are [9]

1. Web and social media data
2. Machine to machine data
3. Biometric data
4. Human generated data

To support the objective, human generated data is collected using a "WhatsApp" a messenger application and through E- mail. An approximate of about 1300 instances was collected from a population group ageing above 18. The above collected data is in unstructured and semi structured format which contribute the Big Data in variety format. The data thus collected has to be transformed and preprocessed.

B. Consolidation

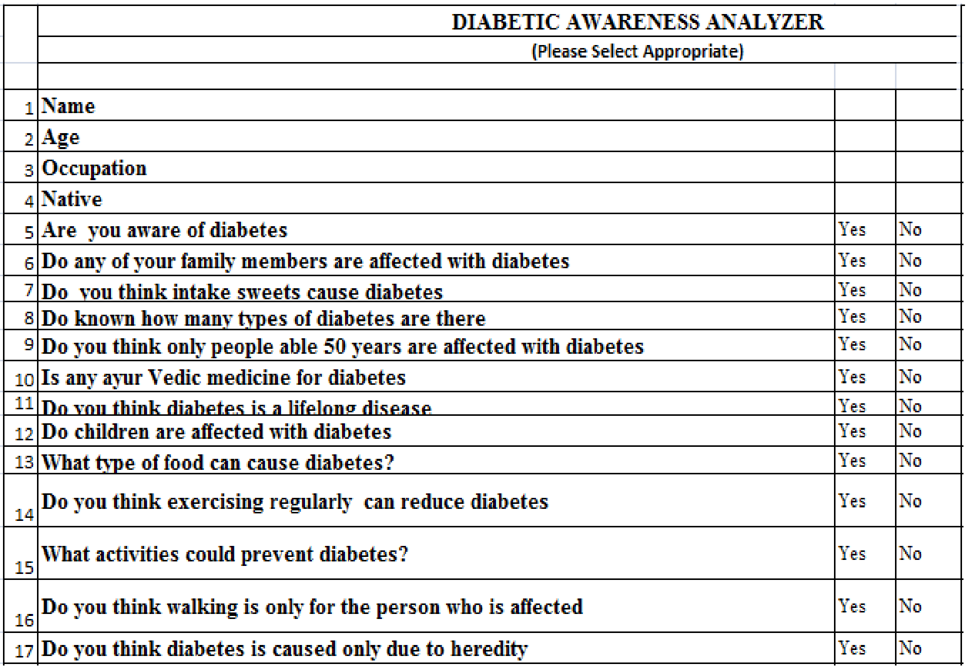
Pre-processing of "WhatsApp" a mobile app data and e-mail data (unstructured) is completely, different from the pre- processing done using KDD process in regular text datasets [10]. Nehal G.Karelia Prof and Shweta Shukla has used web mining algorithm to preprocess the unstructured web blog data[5][11].The data from WhatsApp messenger are like SMS (Short Message Service), the data are in very different slang than common word in English. The text mining approaches are used to preprocess the WhatsApp, E-mail data. Initially the repeated messages are eliminated. As a second step the hash tag, punctuations are removed since it doesn't make any sense and it affects the accuracy of the process [12]. After the removal the messages are tokenized and stemmed. The terms relevant to our study are filtered and stored along with the user information.

C. *Data Analytics using R*

R is a free software programming language. The R language is widely used among statisticians and data miners for developing statistical software and data analysis. The source code for the R software environment is written primarily in C, FORTRAN and R. R uses a command line interface; however, several graphical user interfaces are available for use with R. R provides a wide variety of statistical and graphical techniques, including linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, and others. The consolidated preprocessed data are analysed using R. the statistical methods are used in the exploring analysis of the data. The descriptive statistics and the frequency are used to describe the behavior of the people. The estimation of age standardization prevalence of Knowledge on diabetes, foods that cause and controls diabetes, Symptoms, Occupation, Native are analysed on the data collected by direct standardization methods. Visual and Exploratory data analysis methods are used to find the relation among the response of the people for each questions.

III. RESULTS AND DISCUSSION

The real time data is collected using “WhatsApp” and E- mail. The data is collected from 1300 individuals. The data set consistutes of 19 attributes and 1300 instance like Name, Date of birth, Occupation, Food habits, and some instances related to awareness of diabetes. The data set collected is unstructured format consisting of noise and incomplete data, using pre- processing techniques noisy and inconsistent data are reduced. The regular text mining methods are implied to convert the unstructured data to structured data by the process such as stop word removal, tokenization and stemming. The dataset consisted of many missing values which were replaced using KNN impute method. The figure 2 gives a sample shot of samplequestionspostedbyemailandwhatsApp.Table1gives a sample data collected with the pre-processed data.



	DIABETIC AWARENESS ANALYZER		
	(Please Select Appropriate)		
1	Name		
2	Age		
3	Occupation		
4	Native		
5	Are you aware of diabetes	Yes	No
6	Do any of your family members are affected with diabetes	Yes	No
7	Do you think intake sweets cause diabetes	Yes	No
8	Do known how many types of diabetes are there	Yes	No
9	Do you think only people able 50 years are affected with diabetes	Yes	No
10	Is any ayur Vedic medicine for diabetes	Yes	No
11	Do you think diabetes is a lifelong disease	Yes	No
12	Do children are affected with diabetes	Yes	No
13	What type of food can cause diabetes?	Yes	No
14	Do you think exercising regularly can reduce diabetes	Yes	No
15	What activities could prevent diabetes?	Yes	No
16	Do you think walking is only for the person who is affected	Yes	No
17	Do you think diabetes is caused only due to heredity	Yes	No

Figure 2. Questions posted on Whats App and E-mail

The preprocessed data are loaded into HDFS in structured format. The analysis of the data is done using Visual data exploration, which is an informal way of exploring data. Different plots and graphs are used in Visual Data exploration. Here the plots of awareness among different age group, occupation, native, education are visualized.

The major population responded to the queries are from the Coimbatore district, it is found using the density plot function in R. Figure 3 gives a density curve of people response to the diabetic awareness analyzer. It is found that the mean age of the people responded ranges from 23 to 25 using the R statistical measure. Figure 4 gives a mean age of persons responded to the diabetes analyzer.

Table I. Sample unstructured data, Pre-processed structured data

S.No	**Questions**	**Collected data**	**Pre-processed data**
1	Name	R.Ramya	R.Ramya
2	Age	24	24
3	Occupation	Student	Student
4	Native	Coimbatore	Coimbatore
5	Are you aware of diabetes	Yes	Yes
6	Do any of your family members are affected with diabetes	No	No
7	Do you think intake sweets cause diabetes	No idea	No
8	Do known how many types of diabetes are there	I don't know	No
9	Do you think only people able 50 years are affected with diabetes	No	No
10	Is any ayur Vedic medicine for diabetes	May be	Yes
11	Do you think diabetes is a lifelong disease	Yes	Yes
12	Do children are affected with diabetes	No idea	No
13	What type of food can cause diabetes?	Only based on diet	Diet
14	Do you think exercising regularly can reduce diabetes	Yes, possible by doing	Yes
15	What activities could prevent diabetes?	Exercise, Yoga	Exercise, Yoga
16	Do you think walking is only for the person who is affected	Yes	Yes
17	Do you think diabetes is caused only due to heredity	Yes	Yes
18	Is, there any food items that can reduce diabetes?	Grains	Millets
19	What are the symptoms for diabetes?	Increasing of weight	Increase of weight

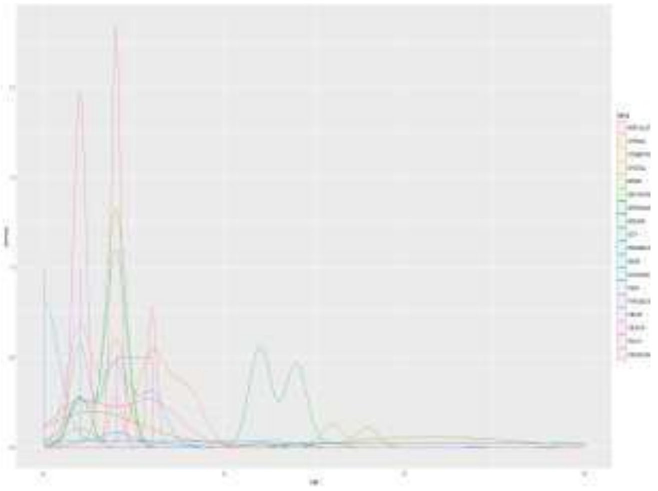


Figure 3. People response to the diabetic awareness analyzer

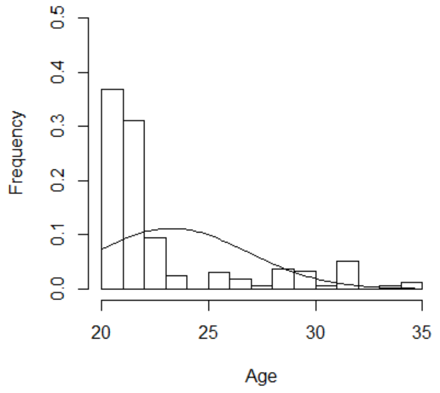


Figure 4. Average age of people response to the diabetic awareness analyzer

From the questionnaire it is found that about 80 % of the people are aware of diabetes, but with deep exploration it is found that a mere of 8.5% of the total population are aware of the types of diabetes as in figure 5.

It's clearly visualized form the figure 7 that above half of the population is not aware of the foods that help in control of diabetes and that helps in reduction of glycemic production in the human body. The figure 8 infers that the people are aware of the symptoms of diabetes as major of the population replied correctly about the symptoms (increased thirst, frequent urination, weight loss, vision problem, etc.,).

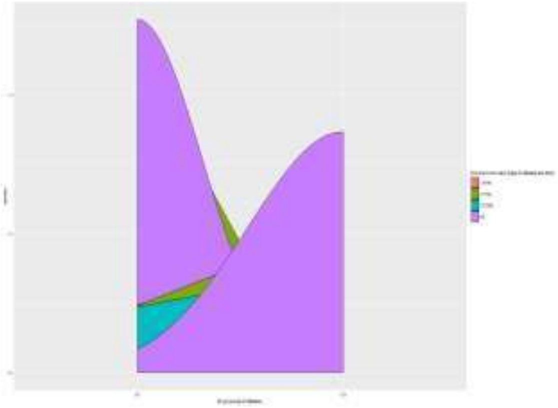


Figure 5. Knowledge of people about the types of diabetes

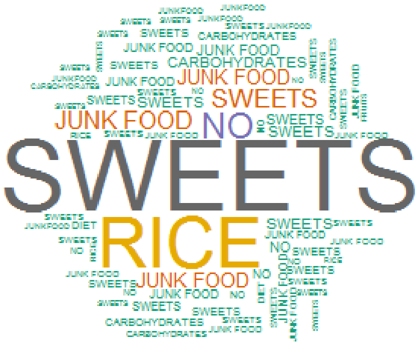


Figure 6. Knowledge of food that causes diabetes

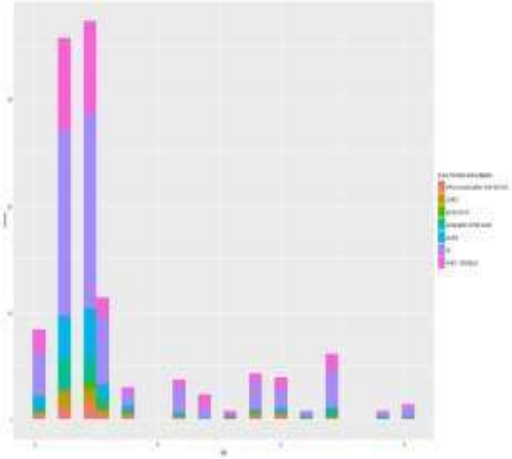


Figure 7. Knowledge of people about foods that help in reduction of diabetes

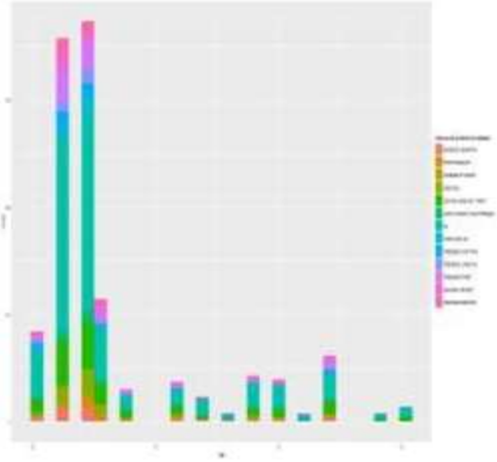


Figure 8. Knowledge of people about symptoms of diabetes

Table II. Details the percentage of people awareness about diabetes

Criteria	**Outcome**
Aware of diabetes	92% of People are not aware about the types of diabetes
Age of occurrence of diabetes	91% are not aware of the age of occurrence of Diabetes.
Food causing Diabetes	80% are aware of foods that cause diabetes
Food items that can reduce diabetes	45% are aware of food items that reduce diabetes, and the reaming are not aware
Symptoms for diabetes	38% are aware about the symptoms of diabetes

IV. CONCLUSION

The awareness of Diabetes mellitus among the population is analysed using R Hadoop. An experimental analysis is made with a Whats App, E-mail data set with 1300 record as a result of the analysis the awareness of the Diabetes among the population is very less is identified. The Analysis is made based on age group, Occupation on the people and the Location they are living. The analysis is made on the real time data, it is found only average on 20-25 Percent of people are aware about Diabetes but they don’t about the proper cause for diabetes mellitus.

REFERENCES

[1] Nawsher Khan, Big Data: Survey, Technologies, Opportunities, and Challenges, Hindawi Publishing Corporation the Scientific World Journal, 2014.

[2] Borthakur D, HDFS architecture guide, hadoop apacheproject., http:/ /hadoop.apache.org /docs /r1.2.1 /hdfs_design.pdf, 2008.

[3] Hadoop. http://hadoop.apache.org/

[4] Big Data Analytics in Health , Canada Health Infoway, 2013.

[5] Raghupathi and Raghupathi, Big Data analytics in health care :promise and potential, Health Information Science and Systems, 2014.

[6] N.M. Saravana Kumar, Predictive Methodology for Diabetic Data Analysis in Big Data, Procedia Computer Science, pp.203 – 208, 2015.

[7] Anne Cooper, Partha Kar, A new dawn: The role of social media in diabetes education, Journal of Diabetes Nursing, pp.68–71, 2014.

[8] International Diabetes Federation, http://www.idf.org/diabetesatlas, 2011.

[9] Ramsingh J and Bhuvaneswari V, An Insight on Big Data Analytics Using Pig Script, International Journal of Emerging Trends & Technology in Computer Science (IJETTCS),Vol.4, No.6, 2015.

[10] Nehal G. Karelia Prof. Shweta Shukla, Data Preprocessing: A Pre requisite for Web Log Files, International Journal of Engineering Research & Technology (IJERT), 2014.

[11] Rohit Pitre, Vijay Kolekar, A Survey Paper on Data Mining With Big Data, 2014.

[12] V.Jude Nirmal and D.I. George Amalarethinam, Parallel Implementation of Big Data Pre-Processing Alorithms for sentimal analysis of Social Networking Data, International journal of fuzzy mathematical archive, Vol.6, No.2, pp.149-159, 2015.