Data Analytic on Diabetic awareness with Hadoop Streaming using Map Reduce in Python

J. Ramsingh
Department of Computer Applications
Bharathiar University
Coimbatore, India

V.Bhuvaneswari
Department of Computer Applications
Bharathiar University
Coimbatore, India

***Abstract*—Thoughts and ideas of the majority of the population are influenced by the opinions and thoughts of the people around them. In this digital era the people are influenced digitally by the comments of other people on social networking sites. Sentiment analysis on those comments would reveal the thoughts of the people. A framework that counts positive and negative opinion about the product is been proposed. The framework makes use of Hadoop Distributed File System (HDFS) to store data set and runs on MapReduce architecture to perform diabetic analysis.**

Keywords*— *MapReduce; Hadoop; Sentiment analysis; HDFS

I. INTRODUCTION

In an era of data deluge, social media has delivered a new standard of interaction between people. The exceptional growth of social media giants ultimately reflects forceful desire to connect with each other around common interests. This presents with a platform for communication and selling products online. In this digital world social media is not just an business tool but works in a greater extent to influence people thoughts and actions. The social networking sites are the main major sources of Big Data.

Big Data is a normal data that are huge in size with lots of information in different format and lots of noise that cannot be mined using the traditional system. Sam Madden [21] stated that the data are too big, too fast, too hard and too complex to analyze with the existing system which is known as Big Data.

The process of storing, analyzing, managing and visualizing the data is very difficult. According to Marko Grobelnik [10] Big Data is very similar to Small-data, Big Data requires a completely new tools and techniques to analyze and solve many real world problems in a better and an efficient way.

Vibhavari in [24] states the Big Data is a most popular word used to denote the massive growth of data in variety of formats(structured and unstructured data), which is hard to be processed using traditional processing systems.

The generation of huge volume of data (Big Data) leads to a development of an analytics called Big Data Analytics. Big Data Analytics is an sophisticated analytic technique used to analyze different types (structured, unstructured and semi-structured) and size of data (terabytes to geophytes). The analytical process is used by many researchers, analysts and business people to make quick and accurate decisions. Big Data Analytics in healthcare contribute a major role in processing and analyzing the data in variety of forms to deliver suitable insights.

Social networking is one of the effective tool to make people aware of a particular product and easy reachable. 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 many health care topics.

India with a population of billion plus people, one of the world's leading growing economy, 29.5% of population are underneath the estimated minimum level of income, 46% of offspring are half-starved. Increasing population in India is a great threat to the health care structure of the country. In India currently 61.3 million people are affected with diabetics, a survey says that peoples in India affected with diabetes will increase to 103 million by 2030 [25] 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. When Diabetes Mellitus is taken in to account many blogs, social networking sites, mobile apps contribute to the generation of Big Data.

According to International Diabetes Federation (IDF) people affected with diabetes build self-management through the resources available in many online blogs, social network communities and patient self-help groups. N.M. Saravana Kumar et al., [12] stated that people with diabetes spending a varying amount of time on self-care, with an average of about 20 minutes per day.

The awareness of diabetic among Indian's are very poor and many are not even diagnosed with diabetics. 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 on Diabetes in India has shown a greater scale of increase in people affected with Diabetes from 8.2% -18.6% in urban areas, and from 2.4%-9.2% in rural areas, with a time frame of just 16 years (1992-2008)[9]. A study of epidemic is necessary to understand the risk factors of Diabetes Mellitus.

The objective of the study is to analyze awareness of diabetes among the people using social media data. The

978-1-5090-3770-4/16/$31.00©2016 IEEE

dataset were collected from social media like tweets and WhatsApp. A diagnostic analytics is carried out to understand the awareness of food preference among the people that cause and cure diabetes and its risk factors. This analytics is performed with a dataset of 1300 instance which comes under Variety (structure and unstructured data) concept in Big Datausing Hadoop streaming in python. The paper is organized as follows the section II discusses about an over view of Big Data technologies. Section III describes a framework for analyze the diabetic awareness among the people using Map reduce algorithm in python. In section IV the experimental results are discussed followed by conclusion in section V.

II. BIGDATA A VIEW

Big Data is a data, available at heterogeneous, autonomous sources, in extreme large amount, which get updated in fractions of seconds [24]. Puneet Singh Duggalet, et. al., [17] stated that Big Data is a heterogeneous mix of both structured and unstructured which required massive storage space and effective methods to manage. Marko Grobelnik [10], Nawsheret, et. al., [13] stated that there would be maximum growth of digital data to 35 trillion gigabytes in 2020. Big Data has become the new front line of today's information management by generating and consuming large amount of system data. which has motivated the need for advanced recent technology from accumulation of data to visualization of results.

Big Data Analytics (BDA) is a new information management approach that has been designed to derive previously untapped intelligence and insights from data [24]. Big Data Analytics is a analytic technique that operates on Big Data sets to make a better analytics. The advanced computing techniques and the analytical techniques can manage the Big Data for better analytics without using super computer or spending high cost for handling them. The advance tools and technology can store, access, and analyze large amounts of data very efficiently with less processing time. The most commonly used tools to handle Big Data are Hadoop and its components [16]. Hadoop [7] is an open source framework used for storing and processing large datasets. The framework is written in java built on the basis of distributed processing system to process large data sets by simple programming model. Hadoop is designed to have high fault tolerance scaling from single server to thousands of nodes [7]. The various components of a Hadoop Stack are HDFS, MapRedue, Hive, HBase, Pig, Zookeeper, Oozietc,.

Borthakur D [5] and Vavilapalli V K et.al., [23] stated that Hadoop Distributed File System (HDFS) (storage),Hadoop YARN(Scheduler) and MapReduce (Processing) are the main three pills of Hadoop Framework.

A. *Hadoop Distributed File System (HDFS)*

The Hadoop Distributed File System (HDFS) is designed based on the Google File System (GFS), the master/slave architecture. The master consists of a single name node and one or more data nodes. The name node manages the metadata and data nodes stores the actual data[13]. The name node decides the mapping of blocks to the data nodes. The data nodes take care of both read and write operation in the file system.

B. *Hadoop Map Reduce*

Hadoop map reduce is a software designed as parallel architecture running on large clusters. The map reduce algorithm consists of two important tasks, map and reduce task. The map task splits the data set into tuples (key/value pairs)[8] and sorts the output and is fed as an input to the reduce task. The input and the output of the job are stored in file system. The process of scheduling, the execution of the failed jobs, monitoring is carried out using the framework.

C. *YARN*

YARN (Yet Another Resource Negotiator) is a cluster management technology, in the second-generation of Hadoop, designed from the experience gained from the First generation. YARN provides a central platform to deliver consistent operations, security, and data governance tools across Hadoop clusters [8] [23].

Hadoop Map Reduce allows various languages to integrate such as Python, C++ etc., [26]. The proposed work is achieved using python a open source scripting language executed in the MapReduce frame work using Hadoop streaming interface , a utility that comes with the Hadoop distribution. This utility allows us to create and run MapReduce jobs with any script as the mapper and the reducer. The mapper reads the data through STDIN a utility in Hadoop streaming [26]and sends the mapped key value pairs to the reducer and the result from the reducer task is stored in the HDFS using STDOUT. The analyzed results are visualized using R- Hadoop. 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.

PiyushGupta et. al., [16] has made use of Hadoop streaming to perform sentiment analysis on user logs. This work makes uses of the Hadoop streaming interface to analyze the awareness of diabetes from the social media data.

III. FRAME WORK AND METHODOLOGY

This section presents with a framework to analyze people awareness on the food that cause diabetes and its risk factors. The framework is implemented using Python on Hadoop streaming.

The framework consists of two phases the Mapper phase and the Reducer phase. The data set is integrated from the data source of "WhatsApp" and twitter data. The questions are posted in WhatsApp and twitter to understand the food preference among global users. A total of 1300 instances was collected from a population group with age above 18, which constitute 60% data from "WhatsApp" and 40% data from Twitter. As the data set collected is in unstructured and semi structured format the data set fit to the character of Variety format. The data thus collected is transformed and pre processed.

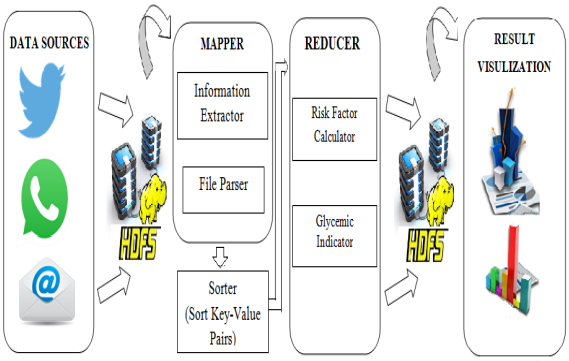

Fig 1. Awareness analyzer

The Fig. 1. represents the framework for the proposed methodology. In the first step the collected data are moved to HDFS. The transformed data set need to be preprocessed, the Pre-processing of collected data (unstructured) is completely, different from the pre-processing done using KDD process in regular text datasets. 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 data.

Initially in the mapping phase the data are read from HDFS using STDIN, the repeated messages are eliminated and the hash tag, punctuations are removed. Stop words in the dataset are identify and removed from the dataset using Stop word dictionary. After the removal the messages are tokenized and stemmed to make the result accurate the diabetic tag are assigned after stemming. Using split method the data is split it into words and output a list of lines mapping words to their (intermediate) counts to STDOUT. The relevant terms to our study are filtered and sorted as a key value pair. The algorithm for mapping function is tabulated in table I.

In the reducing phase the output from the mapping phase is reduced by counting the occurrence of the frequent items. The results are finally stored in HDFS. The R-Hadoop is used to visualize the result. The algorithm for reducing function is tabulated in table II.

The experimental results are discussed in the next section.

TABLE I. PSEUDO CODE - MAPPER

Algorithm : mapper
Get input (standard input)STDIN Import required library for natural language process Load imported library Iterate the input to preprocess: Perform pre preprocessing split each line on whitespace Stem the input word using stemmer Iterate the stemmed word to remove punctuations: Remove punctuation, numbers, stop words, case conversion write the results to STDOUT (standard output); The output here will be the input for the Reduce step, i.e. the input for reducer.py

TABLE II. PSEUDO CODE - REDUCER

Algorithm: reducer.py
Reads input comes from STDIN i.e. output from Mapper (standard input) Initialize the required variables to count the current words Iterate the input to cluster the word: Parse the input we got from mapper.py Convert the count to integer Calculate the count of the words frequency print result to STDOUT store the result in the HDFS

The map reduce task for analyzing the diabetes risk factor is completed in 0.73 milliseconds.

IV. RESULTS AND DISCUSSION

The social media data is collected from 1300 individuals. The data set constitutes of 19 attributes and 1300 instance like Name, Date of birth, Occupation, Food habits, Food that cause diabetes, Food that control diabetes, Symptoms and some instances related to awareness of diabetes.

1	Name
2	Age
3	Occupation
4	Native
5	Are you aware of diabetes
6	Do any of your family members are affected with diabetes
7	Do you think intake sweets cause diabetes
8	Do known how many types of diabetes are there
9	Do you think only people able 50 years are affected with diabetes
10	Is any ayur Vedic medicine for diabetes
11	Do you think diabetes is a lifelong disease
12	Do children are affected with diabetes
13	What type of food can cause diabetes?
14	Do you think exercising regularly can reduce diabetes
15	What activities could prevent diabetes?
16	Do you think walking is only for the person who is affected
17	Do you think diabetes is caused only due to heredity

Fig 2. Sample Data

The Fig. 2. gives a sample shot of sample questions posted in tweets, whatsApp. The data collected are moved to HDFS and in the mapping phase the collected data are preprocessed and the results are sorted as key value pair. In the reducing phase the sorted data is saved into python dictionary and the words are counted based on key value pair. The Figure 3 shows the success full run of map and reduce job on single node cluster and the result stored in the HDFS directory.

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

```
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=2
Job Counters
        Launched map tasks=2
        Launched reduce tasks=1
        Data-local map tasks=2
        Total time spent by all maps in occupied slots (ms)=4732
        Total time spent by all reduces in occupied slots (ms)=2340
        Total time spent by all map tasks (ms)=4732
        Total time spent by all reduce tasks (ms)=2340
        Total vcore-seconds taken by all map tasks=4732
        Total vcore-seconds taken by all reduce tasks=2340
        Total megabyte-seconds taken by all map tasks=4845568
        Total megabyte-seconds taken by all reduce tasks=2396160
Map-Reduce Framework
        Map input records=1200
        Map output records=1200
        Map output bytes=28214
        Map output materialized bytes=30626
        Input split bytes=188
        Combine input records=0
        Combine output records=0
        Reduce input groups=8
        Reduce shuffle bytes=30626
        Reduce input records=1200
        Reduce output records=7
        Spilled Records=2400
        Shuffled Maps =2
        Failed Shuffles=0
        Merged Map outputs=2
        GC time elapsed (ms)=51
        CPU time spent (ms)=1760
        Physical memory (bytes) snapshot=597581824
        Virtual memory (bytes) snapshot=1470455808
        Total committed heap usage (bytes)=511705088
Shuffle Errors
        BAD_ID=0
        CONNECTION=0
        IO_ERROR=0
        WRONG_LENGTH=0
        WRONG_MAP=0
        WRONG_REDUCE=0
File Input Format Counters
        Bytes Read=30993
File Output Format Counters
        Bytes Written=130
6/09/08 19:37:22 INFO streaming.StreamJob: Output directory: /wordcount/my-output.txt
```

Fig . 3. Execution of Mapreduce

A. Experimental Discussions

Based on diagnostic analytics on the dataset the following knowledge are inferred.

1) Knowledge On Food Cause Diabetes

From the Fig. 4. it is analyzed that

- 20% of people thinks that's Sweets are the major cause for diabetes.
- 18% people opinioned that food rich in carbohydrate may lead to diabetes.
- 2% opinioned intake of Rice leads to diabetes
- 10% felt Junk foods, Non- Veg, and oil food may leads to diabetes.
- 50% of people had no idea about the foods that lead to diabetes.

Fig 4.Word Cloud for analyzing cause Diabetes

2) Knowledge On Food Controls Diabetes

From the word cloud in figure 5, It clear inferred that

- 8 % of people opinioned Tiredness and loss of wait is the symptoms of diabetes.
- 10% people felt frequent urination is the symptoms for diabetes.
- 2% people opinioned Headache, Foot Pain and weakness are the symptoms for diabetes.
- 80% of people had no idea for the symptoms of diabetes.

Fig 5.Word Cloud for analyzing Symptoms that cause Diabetes

3) Knowledge On Symptoms For Diabetes

From word cloud in Fig. 6. it is clearly found that

- 15% of people think that wheat can control diabetes.
- 15% of people opinioned that fruits and vegetables will control diabetes.
- 10% of people suggested millets, fish and nuts can control diabetes.
- 70% of people are unaware of the food items that control diabetes.

Fig 6.Word Cloud for analyzing food that control Diabetes

The table III details the percentage of people awareness about diabetes. Fig. 7. represents the diabetic awareness among people based on food items that can cause and control diabetics and the symptoms of diabetes.

TABLE III.OUTCOME OF THE ANALYSIS

Criteria	Outcome
Food causing Diabetes	50% are not aware of foods that cause diabetes
Food items that can reduce diabetes	40% are aware of food items that reduce diabetes, and the reaming are not aware
Symptoms for diabetes	20% are aware about the symptoms of diabetes

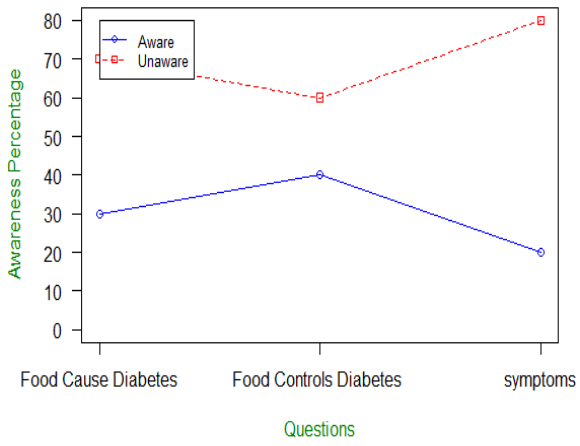

Fig 7. Diabetic awareness comparison

V. CONCLUSION

An experimental analysis is made using social media data collected from tweets and WhatsApp data to analyze the awareness of Diabetes mellitus among the people based on their views. A methodology for analyzing the social media data using MapReduce is implemented and the results are analyzed. From the analytics it is found that only an average less than 50 Percent of people are aware about the food items that may lead to Diabetes mellitus and people less than 10 percent are aware about symptoms of diabetes. It is also found that the awareness is high among the age group from 20 to 25. The overall analyze is concluded that the people are less aware about diabetes mellitus, and there is an urge to make people aware and prevent diabetes. The sad part is that the growing younger generations are more accountable to diabetes.

References

[1]. Anne Cooper, ParthaKar, "A new dawn: The role of social media in diabetes education", Journal of Diabetes Nursing 18: 68–71, 2014.

[2]. Asmi, A., & Ishaya, T. "Negation identification and calculation in sentiment analysis." IMMM 2012, The Second International Conference on Advances in Information Mining and Management. 2012.

[3]. Batool, R.; Khattak, A.M.; Maqbool, J.; Sungyoung Lee, "Precise tweet classification and sentiment analysis," Computer and Information Science (ICIS), 2013 IEEE/ACIS 12th International Conference on , vol., no., pp.461,466, 16-20 June 2013.

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

[5]. Borthakur D, "HDFS architecture guide" (2008). HADOOPAPACHEPROJECT.http://hadoop.apache.org/docs/r1.2.1/hdfs_design.pdf

[6]. Godbole, N., Srinivasaiah, M., &Skiena, S. (2007).“Large-Scale Sentiment Analysis for News and Blogs." ICWSM 7 2007 .

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

[8]. Interface Mapper, (Last visited in June 2015). [online].Available:http://hadoop.apache.org/docs/r2.6.0/api/org/apache/hadoop/mapred/Mapper.html

[9]. International Diabetes Federation. Diabetes atlas. 6th edn.Brussels: IDF, 2011. http://www.idf.org/diabetesatlas - accessed14 February 2014.

[10]. Marko grobelnik " Big Data tutorial " jozef stefan institute ljubljana, slovenia http://www.planet-data.eu/sites/default/files/presentations/Big_Data_Tutorial_part4.pdf

[11]. Minqing Hu and Bing Liu. "Mining and Summarizing Customer Reviews." Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004), Aug 22-25, 2004, Seattle, Washington, USA.

[12]. N.M. Saravana Kumar et al., "Predictive Methodology for Diabetic Data Analysis in Big Data", Procedia Computer Science 50, 203 – 208, 2015.

[13]. Nawsher Khan et al "Big Data: Survey, Technologies,Opportunities, and Challenges", Hindawi Publishing Corporationthe Scientific World Journal Volume 2014, Article ID 712826

[14]. Nehal G. Karelia Prof. Shweta Shukla, "Data Preprocessing: A Pre requisite for Web Log Files", International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181.

[15]. Opinion Mining, Sentiment Analysis, and Opinion Spam Detection, (Last visited in june 2015)

[16]. Piyushguptha et al,”Sentiment analysis on hadoop with hadoopstreaming”IJCA, volume 121-no 1.1,July 2015.

[17]. Puneet Singh Duggal, "Big Data Analysis: Challenges andSolutions", November, 2013.

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

[19]. Ramsingh J and Bhuvaneswari V, "An Insight on Big DataAnalytics Using Pig Script ", International Journal of EmergingTrends & Technology in Computer Science (IJETTCS),Volume 4,Issue 6, November - December 2015 ISSN 2278-6856.

[20]. RohitPitre, Vijay Kolekar, "A Survey Paper on Data Mining With Big Data"

[21]. Sam Madden "From databases to Big Data" ieee computer society 1089-7801 2012.

[22]. stop-words, (Last visited in june 2015) [online]. Availaible: https://code.google.com/p/stop-words/

[23]. Vavilapalli VK, Murthy AC, Douglas C, Agarwal S, KonarM,Evans R, Graves T, Lowe J, Shah H, Seth S, "Apache hadoopyarn: Yet another resource negotiator", In: Proceedings of the 4thannual Symposium on Cloud Computing., p 5, 2013.

[24]. Vibhavari Chavan et al, "survey paper on Big Data" (IJCSIT) International Journal of Computer Science And Information Technologies, vol. 5 (6) , 2014, 7932-7939

[25]. World Health Organization, Regional Office for South East Asia.Health situation in the South-East Asia Region 1998-2000. NewDelhi: WHO-SEARO, 2002. Document No. SEA/HS/222.http://209.61.208.233/LinkFiles/Health_Situation_toc+forward.pdf

[26]. Writing Hadoop Applications in Python, (Last visited in July 2016). [Online]. Available: http://www.glennklockwood.com/di/hadoop-streaming.php

[27]. Xiaoqian Zhang; Shoushan Li; Guodong Zhou; Hongxia Zhao, "Polarity Shifting: Corpus Construction and Analysis, "Asian Language Processing (IALP)", 2011 International Conference, pp.272-275, 15-17 Nov.2011.