International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-2S2, January 2019

Prediction of Active Users and Location Surveillance to Study Diabetes Disease Dynamics

P.Vasudha Rani, K.Sandhya Rani

Abstract - Twitter is the source for a massive amount of spatiotemporal information about individuals propagating their opinions, feelings, suggestions, support, etc. These data can be utilized in a number of diverse fields to monitor and understand a range of social phenomena. One important and largely unexploited use of Twitter's massive data feed is to utilize tweets as a means for understanding trends in public health. Twitter-based health research is a growing field funded by a diversity of organizations. As a part of the health-related research, here the task is to analyze Diabetes related User Activity and Location-specific postings to identify the Active Locations with more Active Users that generates more user Activity. This task requires geolocated tweets with user details and other information such as followers and follows.

As a part of the health-related research, a model is proposed in this paper which focus on tracking and understanding the patterns of causes, affected diseases and suggested foods for the control and prevention of Diabetes in a comprehensive approach. For this work, the data source is a tweet source generated from individual and majority group twitter accounts working on Diabetes and related issues continuously. The proposed method uses an advanced data mining framework with a novel use of social media data retrieval and classification filtering to identify Active Locations & Users and understand how tweets from Active Users and Active Locations can be used to explore the prevalence of causes, affected other diseases, and healthy food suggestions for control and prevention of Diabetes.

Keywords - Active Users, Active Locations, Classification, Geolocated Diabetes Tweets, Patterns of Causes, patterns of Affected Diseases.

I. INTRODUCTION

The main objective of this paper is to track and analyze Diabetes Tweets data to predict Diabetes disease scenarios from the posts of twitter. With the vast amount of available twitter data, the ability of researchers in handling Big data is increasing in a broader length. It is very much important to explore the potential of usage of this data in a way that it gives prior awareness on health-related aspects to the public. One of the three main functions of public health is the "assessment and monitoring of the health of communities and populations at risk to identify health problems and priorities.

For decades, health researchers have leveraged large databases of health information for this purpose. In recent years, researchers have recognized that social media platforms, such as Twitter, Facebook, and Instagram, can also provide data about population-level health and behavior. And it is also possible to trace any new symptom or event related to a disease in tweets too early before and be traced with the reports of the Centre for Disease Control. According to an International Data Corporation (IDC) report sponsored by Seagate Technology, it is found that big data is projected to grow faster in healthcare than in sectors like manufacturing, financial services or media. It is estimated that the healthcare data will experience a compound annual growth rate (CAGR) of 36 percent by the year 2025.

The vast network of healthcare influencers, thought-leaders, patients, providers, organizations, and governmental entities daily create rich healthcare content, messages and signals that provide incredible value if it is segmented, analyzed and curated in a meaningful way to answer the unique questions and needs of the public. Among social media networks, Twitter [1] provides a unique big data source for public health researchers because of the real-time nature of the content, and the ease in accessing and searching publically available information. The reach and volume of data are also significant every day, 500 million tweets are sent by more than 300 million active users worldwide.

In addition to its potential as a more traditional data source, Twitter is also interactive, researchers can contribute to the social network and harness this feature as a recruitment tool or for intervention. Despite the potential for this social media platform, the landscape of how Twitter is and might be used for health research has yet to be defined. There is value in understanding the ways that the Twitter [2] data set can be harnessed to contribute to our understanding of public health.

So, here my model is designed to analyze a health issue which is more haunting nowadays i.e. Prediabetes & Diabetes Awareness.

Here the proposed work is regarding Twitter data analysis on "Geolocated Diabetes data set" to provide awareness on Active Locations – From where regular Diabetes related tweet updates into Twitter, Active Users – either an individual Twitter user or a group account user who provides continues diabetes postings into the twitter, and Popular User Accounts- with significant users who have more followers.

All the users and locations data is visualized through Google maps. Only the Active locations and Active Users data is visualized through Network diagrams in R. The user

Manuscript published on 30 January 2019.
* Correspondence Author (s)
P. Vasudha Rani, Research Scholar, CS Dept, SPMVV, Tirupati, Sr. Asst Professor, IT Dept, GMRIT, AP, India
Dr. K. Sandhya Rani, Professor, CS Dept, SPMVV, Tirupati, AP, India

activity of these users is observed using a line plot to find the growth rate of Activity. Finally, the event of identifying popular users by the parameters of tweet frequency, followers and follows counts is visualized through histograms. The application of this is to collect the latest and useful tweets information from the active users [3] and active location for text classification [4] to identify patterns of unigrams, bi-grams, trigrams related to the causes, other diseases affected, suggested foods to visualize in a word cloud.

The above mentioned broad range of prediction applications, social network analysis, and behavioral modeling applications uses twitter data. The theme of the data set is about Prediabetes and Diabetes.

Our aim is not only to retrieve user posted tweets on Prediabetes & Diabetes, but also the important details such as "date of tweet", "user's screen name"," followers", "follows", and lastly "location" –the country from where the tweet was posted. All these attributes play a key role in our research. To describe each attribute , firstly "date of tweet" has got its significance to monitor the growth of activity of a user in tweeting for a period of time, secondly "user's screen names can be utilized to identify and monitor active user's tweets based on their screen names, thirdly "followers" attribute is used to analyze the popularity of the user accounts for diabetes, finally location is a highly significant attribute to identify the top country locations from where more Diabetes updates are coming continuously.

Another important factor is about the classification of twitter accounts. There are two types of twitter accounts basically. They are Individual and Group account. An individual is a type of regular account wherein a group account is maintained by an organization. The motto behind the group account is either an employee supportive web blog or a Health Related NGO. It means multiple people such as nutrients of a company as a single team manages a group account as a single member. Another difference could be that group account has more no of followers. And the majority of twitter group accounts work for an objective.

Another two major focused parameters of this research work are i) Twitter Active users ii) Geolocation Inference [5]. Twitter users are the one who has a Twitter account. Twitter users post their opinions on the Twitter web site about their experiences, sentiments towards any public related events, health-related, diseases related, etc. But Twitter user's vary in their level of activity i.e about the no of tweets they post every day and the regularity in sharing their opinions on Twitter. Twitter user's with a high-end level of activity is called "Active Users".

Active users are characterized by the element of completeness of the tweet which includes the details like location, time stamp, mentioning the other user's details in the tweet which talks about their connected responsibility[6] towards the outside world. More contributions to Twitter content is from active users. Active users can be identified either by their level of activity or by the follower's count. In 2018 end of the year survey, total no of monthly Active users are 326 million and total no of tweets sent per day is 500 million.

Inferring the representative geographic location of social media users is a fast developing area of research with wide applicability in both applied and basic research endeavors that use social media data. Social media based early detection and prediction studies of seasonal flu surveillance [7] or the prediction of commercial movie success require knowledge of user's location.

In this paper, the main focus is to study the Diabetes disease dynamics by predicting active users and surveillance of active locations. A brief description of research related to proposed work are presented in next section.

II. LITERATURE SURVEY

A. Related Works

In 2015, Patrick Park. et al. [3] developed a framework that analyzes the level of activity by different Twitter Users and inferring the representative geographic location of social media users is an actively developing area of research with wide applicability in both applied and basic research endeavors that use social media data. The state-of-the-art location inference method based on label propagation has been shown to perform with high coverage and accuracy. This method relies on the fact that the majority of communication partners, or network neighbors, in the Twitter users mentioned network are geographically proximate.

In 2017, Lauren Sinnenberg BA. et al. [1] have discussed how " Twitter as a Tool for Health Research: A Systematic Review". The work utilized more than 5 billion tweets for the analysis and found that the majority of articles 57% focused on analyzing the content of tweets, whereas other studies harnessed Twitter's interactive features for recruitment or Interventions. Most studies were published in the last 2 years. Twitter-based public health research is a growing field.

In 2013, David A. Broniatowski. et al. [7] have summarized a recently developed influenza infection detection algorithm that automatically distinguishes relevant tweets from other chatter, and we describe our current influenza surveillance system which was actively developed during the full 2012-2013 influenza season.

In 2013, Shaomei Wu. et al. [11] have proposed a model to investigate the classification of users. Based on the classification, we find a striking concentration of attention on Twitter, in that roughly 50% of URL's consumed are generated by just 20K elite users and found significant homophily within categories: celebrities listen to celebrities, while bloggers listen to bloggers.

In 2003, Micheal J. et al. [2] proposed a model that prefers mining of social media outlet Twitter for geolocated messages which provides a rich database of information on people's thoughts and sentiments about numerous topics, like public health. The author utilizes an advanced data mining framework with a novel use of social media data retrieval and sentiment analysis [12-13] to understand how geolocated tweets can be used to explore the prevalence of healthy and unhealthy food across the contiguous United States.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

III. PROPOSED METHODOLOGY

A.Overall System Architecture

The proposed model can efficiently build a framework to work with a health-related issue such as Diabetes & Prediabetes. This framework deals with issues such as exploring the active locations & active users in a social activity or prediction & surveillance of growth rate and popularity of user activity or analyze the negative tweets for the identification of patterns of causes and affected diseases and suggested food patterns from the positive tweets.

The detailed architecture is shown in Fig.1. The proposed model consists of Four levels of work namely i) Data Preparation Phase ii) Identification Phase iii) Exploration Phase iv) Analyze Phase. Each phase involves some subtasks in it. The complete step by step process is demonstrated in the Algorithm1. A detailed description of each phase is given below.

Algorithm 1 Identification of Diabetes Disease Patterns

Input: Tweets Dataset on Diabetes TD

Output: Patterns of Causes, Affected Diseases, Suggested Foods

Classification through Filtering

Step1: Preprocessing TD to filter unnecessary data attributes

Step2: Preprocessing Tweet text to remove unrequired symbols like @,# etc and stop words

Step3: Result is a .csv file with user's Screen Name, date, location, tweet text, followers count, follows count,

Step4: Identification phase-Active locations and Users

Step5: Exploration Phase-Popularity & Growth rate of User Activity

Step6: Analysis Phase-SVM Classification to get Negative and Positive Tweets of Diabetes

Step7: Analysis of Negative Tweets – to design
Word clouds for Causes identified for Diabetes
Word cloud for Affected Diseases of Diabetes

Step8: Analysis of Positive Tweets- to design
Word Cloud for suggested Food patterns

Step9: End

Phase 1: Data Preparation Phase

Data Preparation Phase deals with handling and preparing the Diabetes Tweets data for the next phases of work. This phase involves the process steps as retrieving

Diabetes tweets by executing codes through the Twitter API. After getting Diabetes tweets data set, it is required to remove data from the .csv file such as retweet count, device source, etc. The input .csv file should have the only attributes User's Screen Name, Date of Tweet, Location, Tweet text, No of Followers and Follows count.

After removing unnecessary fields, it is required to remove the duplicate entries in the file. Another key process of this phase is preprocessing the tweet text to remove unnecessary symbols and stop words and etc.

Now the input file is suitable for the active locations and active users identification process with only the required fileds and only the required part of tweet text. This is given as input for the next phase called as Identification Phase.

Phase 2: Identification Phase

The second phase of the work is named as Identification Phase that deals with the execution of the two tasks namely i) Finding Active Locations of Diabetes ii) Finding Active Users of Diabetes. To start with Active Locations Identification, an active location for Diabetes with respect to Twitter is defined as "the location(country) from where people give their updated tweets regularly and completely regarding Diabetes". In this paper, to find the Active country locations, 1 lakh tweets for a period of three months on Diabetes have been gathered with corresponding locations. Now, these Tweets are classified with different labels for multiple times to achieve finally Locations and the corresponding Tweet count.

First classification based on the Location name to find its screen name also. So the generated table has the fields only location name and the screen names.

Next level classification based on Location to map to its frequency count using table command in R. Which in turn cumulates to find the tweet count for each location. Then sorting locations in descending order to get the top active locations [8-9] on the top. Find only the subset of locations whose frequency >=100 to get only the top ones. And the top location countries identified are USA, UK, India, Canada, Australia, Belgium, France, Germany, Singapore.

In the second task of Finding Active Users that classification is done based on the screenname and then executing table command to find the cumulative frequency count for each screenname.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

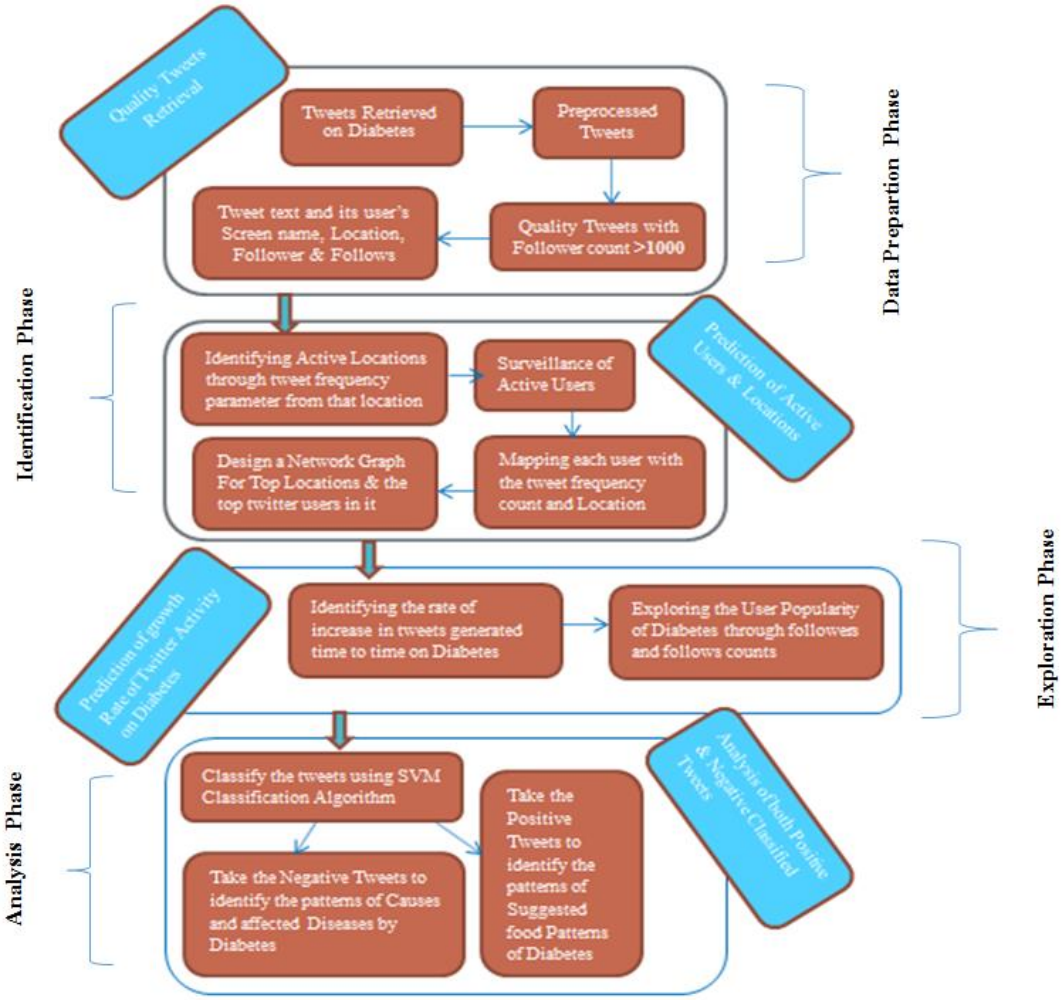


Fig.1. Overall Architecture

Phase 3: Exploration Phase

This is the third key phase of the overall system named as Exploration Phase. The main goal of this phase is to explore some other performance evaluation parameters to perform the popularity analysis of the active users identified in the previous phase. They are "followers count" and "follows count" with the "date of the tweet". This task classifies the tweets data to get a data frame with only the fields "screenname", "followers", "follows". Next, the table command is used to find the frequency counts for each user and then arrange the data in descending order to get the Top users with followers and follows on the top. Get the subset of top tweets using subset command in R. Visualizing the popularity of users is done through bar plot.

The second part of this task is the Prediction of the growth rate of twitter user activity. Consider an active user, for his activity of tweeting, for a period of time, taking the lowest no of tweets and highest no to calculate growth rate of user twitter activity regarding Diabetes. The growth rate of user activity is visualized through line charts.

$$\text{Growth rate} = \frac{(\text{high-low})}{\text{low}} \otimes 100 \quad \text{(i)}$$

(or)

$$\text{Growth rate} = \frac{(\text{no of tweets-no of days})}{\text{no of days considered}} \otimes 100 \quad \text{(ii)}$$

Phase 4: Analysis Phase

This is the last and important phase of the total process. This phase requires Diabetes tweets data that is classified through SVM Sentiment Analysis to derive positive & negative tweets separately. The comprehensive SVM sentiment analysis is discussed in my earlier paper titled "Identification of Ontologies of Prediabetes through SVM Sentiment Analysis". This is a prerequisite process for the analysis phase. The resultant classified Diabetes tweets are the input for analysis phase.

The tweets data set identified with semantic polarities will be the input for the current phase of analysis tasks i.e i) Identifying Causes of Diabetes ii) Finding other effected diseases of Diabetes and Foods suggested for Diabetes. Classified Negative Diabetes Tweets are the Input for analyzing Task-(i) and Positive Tweets are the input for analizing Task-(ii). This phase deals with the main two key analysis tasks i.e i) Analyzing the Negative Tweets ii) Analyzing the Positive Tweets.

Tokenization: Tokenization is the process of dividing a string into substrings based on a property such as unigram, bigram or trigram. All the

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

preprocessed tweet text has been passed through tokenization.

For analyzing the top keywords, using the count based technique to know the no of occurrences of a given unigram, bigram and trigram in our data set. After analysis, we found that people use various vocabularies talking about causes of Diabetes and other diseases affected by Diabetes. First level filtering is based on a threshold factor reaches a minimum count. The second level, it is final verification manually to separate causes of Diabetes and other diseases of diabetes from the remaining keywords. And the keywords are

Causes={"Hypertension","Obesity","Stress","High Cholesterol","Diet genes","Sleep Disturbances","Blood vessel damage","Parental Fat Diet","Human Alzheimer's"}

AffectedDiseases={"Arthritis","Infertility","Autism","Nephrology","Osteoarthritis","Birth defects","Foot ulcers","Heart disease","Heart stroke","Heart Failure","Kidney failure","Organ Transplants","Muscle weakness","Life threatening issue","Cardio Vascular risk","Asthma respiratory disease","Chemic heart disease","High blood pressure","Health imbalances","Memory Loss Dementia"}

The above said tokenization process is repeated for positive tweets to find the suggested food patterns. And they are

Foods Suggested={"Mushrooms","Garlic","Allicin","coffee", "Herbal remedies","Personalized diet","Salad Vegetable","Sea weeds","Turmeric Curcumin","Low carb foods","Plant based meal","Plant based nutrition"}.

All the keywords and their frequencies are stored in an Excel file which is the input for Word Cloud generation to observe the intensities of all the word grams i.e unigrams, bigrams and trigrams related to Diabetes.

IV. EXPERIMENTAL RESULTS

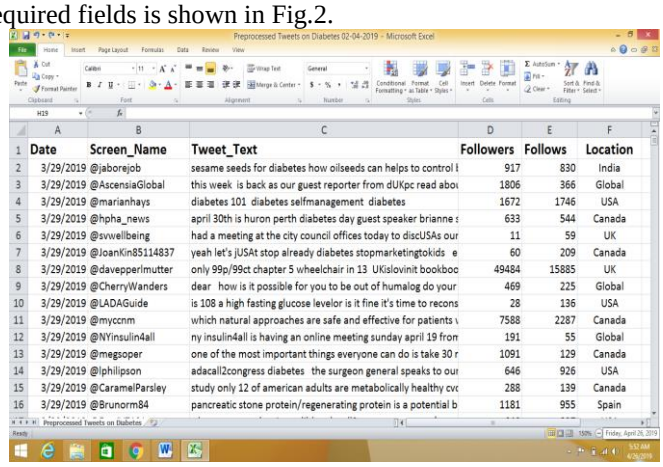
In this work, diabetes Tweets Data is considered for experimental purpose. In addition to tweet text data analysis, time series analysis is also performed to see the Real-time response from users is a major outcome of this work. The time series analysis makes use of other twitter data such as “date of tweet”, “location”, ” followers”, ” follows” etc. These findings are more important because it provides justification for the analysis. The complete experimentation is done using the R framework. The four major tasks identified for implementation are i) Data Preparation ii) Identification iii) Exploration iv) Analysis.

Data Preparation Tasks

Making the Tweets data ready for the process steps, it has to pass through the preprocessing phase. There are two operations to perform i) Remove unnecessary columns ii) Name the columns as per visualization Requirement. The R code for Preprocessing Tweets is given below.

```
#reading the .csv file
MyData <- read.csv(file="Tweets on diabetes 11-04-2019.csv", header=TRUE, sep=",", stringsAsFactors=FALSE)
#removing unnecessary columns and rows by modifying the below statement
MyData <- MyData[-1,]
#giving column names
names(MyData)[1] <- "Date"
names(MyData)[2] <- "Screen_Name"
names(MyData)[3] <- "Tweet_Text"
names(MyData)[4] <- "Followers"
names(MyData)[5] <- "Follows"
names(MyData)[6] <- "Location"
```

After Preprocessing the resultant tweets data only with the required fields is shown in Fig.2.



	A	B	C	D	E	F
1	Date	Screen_Name	Tweet_Text	Followers	Follows	Location
2	3/29/2019	@jaborejob	sesame seeds for diabetes how oilseeds can helps to control l	917	830	India
3	3/29/2019	@AscensiaGlobal	this week is back as our guest reporter from dUKpc read abou	1806	366	Global
4	3/29/2019	@marianhays	diabetes 101 diabetes selfmanagement diabetes	1672	1746	USA
5	3/29/2019	@hpha_news	april 30th is huron perth diabetes day guest speaker brianne s	633	544	Canada
6	3/29/2019	@svwellbeing	had a meeting at the city council offices today to discUSAs our	11	59	UK
7	3/29/2019	@JoanKin85114837	yeah let's jUSAt stop already diabetes stopmarketingtokids e	60	209	Canada
8	3/29/2019	@davepperlmutter	only 99p/99ct chapter 5 wheelchair in 13 UKislovinit bookboo	49484	15885	UK
9	3/29/2019	@CherryWanders	dear how is it possible for you to be out of humalog do your	469	225	Global
10	3/29/2019	@LADAGuide	is 108 a high fasting glucose levelor is it fine it's time to recons	28	136	USA
11	3/29/2019	@myccnm	which natural approaches are safe and effective for patients v	7588	2287	Canada
12	3/29/2019	@NYinsulin4all	ny insulin4all is having an online meeting sunday april 19 fron	191	55	Global
13	3/29/2019	@megsoper	one of the most important things everyone can do is take 30 r	1091	129	Canada
14	3/29/2019	@lphilipson	adacall2congress diabetes the surgeon general speaks to our	646	926	USA
15	3/29/2019	@CaramelParsley	study only 12 of american adults are metabolically healthy cvc	288	139	Canada
16	3/29/2019	@Brunorm84	pancreatic stone protein/regenerating protein is a potential b	1181	955	Spain

Fig.2. Filtered Diabetes Tweets Data with Required Fields

The figure showed that the locations column of the Diabetes Tweets is from different countries such as USA, UK, Canada, Australia, India, Berlin, CA, WA, Nepal, Chicago, Dubai, etc. These Tweet locations of Diabetes can be located in a Google Map. From the Tweet file, we have the location Name only. To find the values of Geolocation attributes “Longitude”, “Lattitude” for each Location we need to register with Google to get an API key from Google. After that Using the command geocode, we can get loc attribute which gives both latitude and longitude. Google map is shown in Fig.3.

```
loc = geocode(locations)
#longitudes   xlon <- loc$lon
#latitudes    ylat <- loc$lat
```

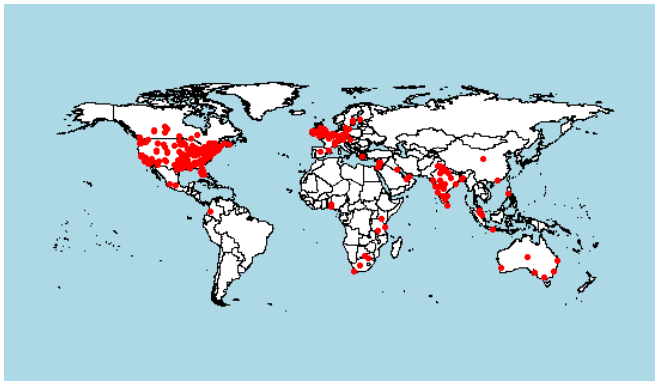


Fig.3. Google Map located with Diabetes Tweets Located Countries

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

B. Identification Phase tasks

This .csv file named as Preprocessed Tweets on Diabetes.csv is the input for the tasks of Identification Phase. They are i) Finding Active Locations of Diabetes ii) Finding Active Users of Diabetes.

1.Finding Active Locations

To predict active locations, the data file shown in Fig.2. is the data set considered for classification filtering. Initially, to map locations and user's screen names, a data frame is created. Then table command is used to cumulate no of entries for each location. That gives the tweet frequency from each country location. Arrange them in descending order to find the top twitter active country locations. The logic behind the task is given below.

```
#reading the preprocessed file
MyData <- read.csv(file="Preprocessed Tweets on Diabetes 02-04-2019.csv",header = TRUE, sep=",",stringsAsFactors = FALSE)
#getting top locations
A <- data.frame(MyData$Location,MyData$Screen_Name)
x <- table(A$MyData.Location)
x <- as.data.frame(x)
x <- x[order(x$Freq,decreasing = TRUE),]
x <- subset(x,x$Freq>=100)
x <- x[-1,]
#getting top users from 6 locations and storing them in a new dataframe
z <- data.frame()
for (i in 1:6) {
B <- subset(A,A$MyData.Location==x[i,1])
y <- table(B$MyData.Screen_Name)
y <- as.data.frame(y)
y <- y[order(y$Freq,decreasing = TRUE),]
y <- data.frame(x[i,1],y[1:10,])
z <- rbind(z,y) }
```

R code is written for identifying top country locations whose tweet count is at least 100. The resultant top counries data is shown in Table 1. For the top 10 active country locations which are the leading countries based on tweet count are visually shown in Fig.4.

Table 1. Top Active Locations with Tweet Count

Location	Tweet Count
Singapore	100
Germany	151
France	188
Belgium	404
Australia	556
Canada	1262
India	2482
UK	4798
Global	10971
USA	17782

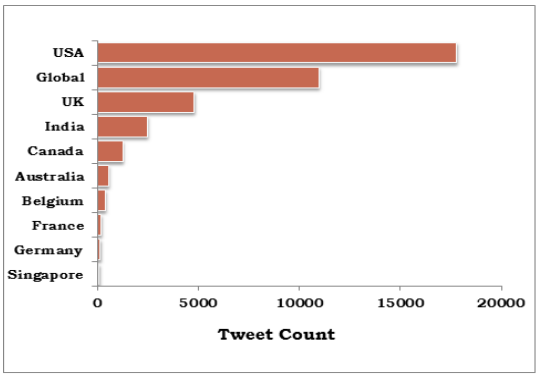


Fig.4. Top Active Countries with their Active Scores of Diabetes Tweets

The execution of the R code results into each location and the corresponding screen users identified. Here we have the Network Diagram Analysis shown for each country. Importantly for the countries listed USA, UK, INDIA. In the data set, some of the location fields are mentioned with country names, others are city names of the country. All the city locations are mapped to its country. It is shown in the Fig.5. for USA and it's identified user's screen names. For all the top ten countries detailed mapping is shown in Fig.6 and Fig.7.

Fig.5. The Country USA and its User ScreenNames

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

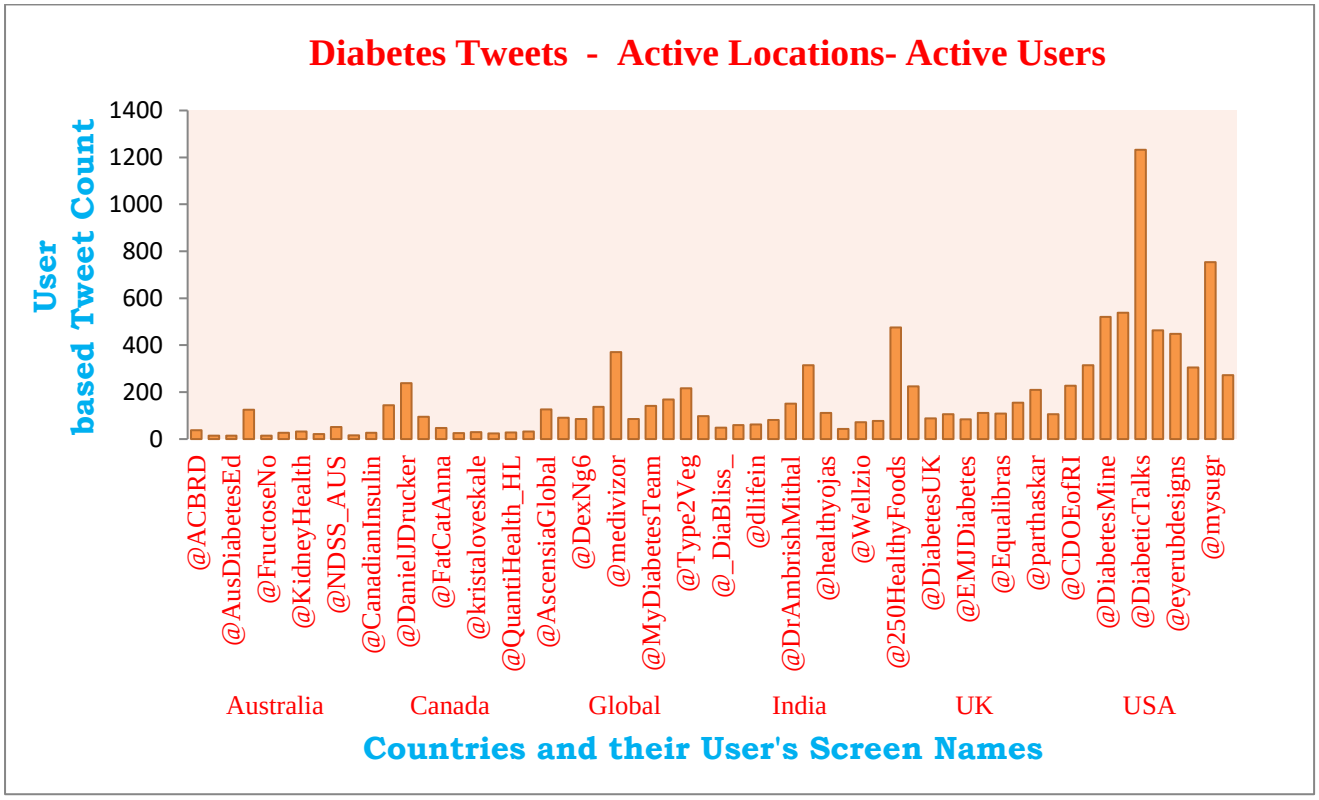


Fig.6. Active Country Locations with the Corresponding Users and their Tweet Frequency Counts

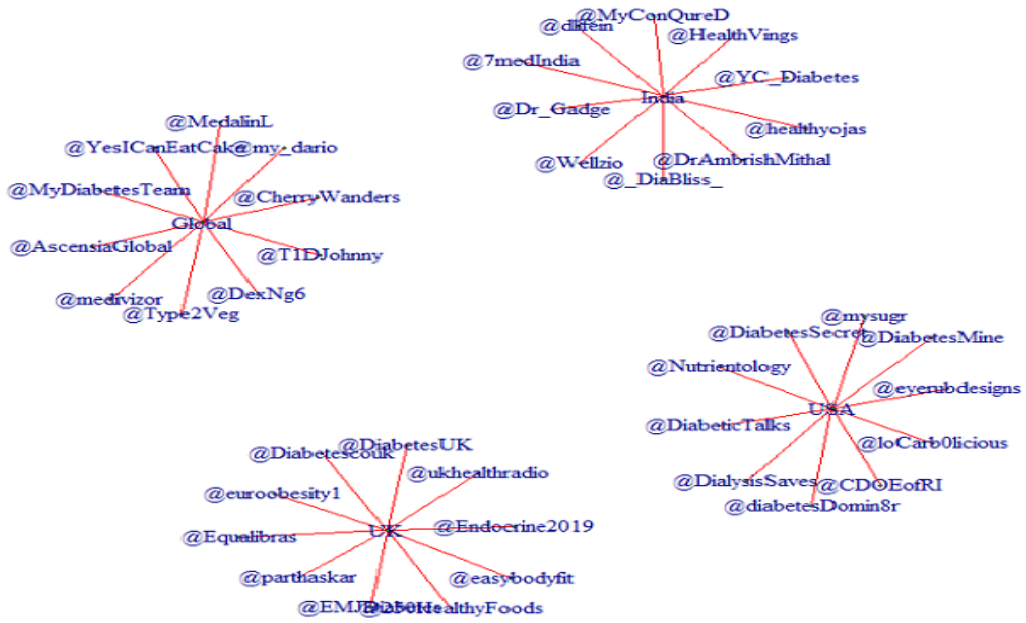


Fig.7. Network Diagram Analysis- Active Countries and the Corresponding User's Screen Name

2. Finding Active Users

Active Users of Twitter for Diabetes are the people who regularly tweet to make people aware of the updates regarding. NGO's working in Global Health Research, are the nongovernmental organizations work for society. health-related NGO's provide valuable resources, tools, and funding in the field of health research. Some health-related NGO's maintain a twitter group account which works with a helping motto. For example, @HealthVings is a health-related twitter group account works with a motto: KEEPING YOU HEALTHY TODAY FOR TOMORROW. So identifying these type of active health Twitter groups definitely help people to know day to day information. The main focus of this task is to identify active users for Diabetes. The classification filtering starts with the preprocessed tweets file with five fields among which date, user's screen name are considered. Execution starts with the mapping of date, screenname into one data frame, so that on each date whether a particular user tweeted or

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

not. Table command in R is used to find the day wise tweet count for each user. Line Chart is used to visualize all the user's activity.

```
#getting top Active Users
C <- data.frame(MyData$Date,MyData$Screen_Name)
C <- C[order(nrow(C):1),] #invert row order
p <- table(C$MyData.Screen_Name)
p <- as.data.frame(p)
p <- p[order(p$Freq,decreasing = TRUE),]
p <- subset(p,p$Freq>=100)
p <- p[1:10,]
```

The execution of this code gives the top users with their tweet counts. And is shown in Table.2. Table data is visualized as a line diagram in Fig.8. & comparison of four users is shown in Fig.9.

Table 2. Top Active Users with their Tweet Count and Country

	User Screen Name	Tweets Count	Country Name
4556	@DiabeticTalks	1233	USA
12275	@mysugr	754	USA
4489	@DiabetesSecret	538	USA
4468	@DiabetesMine	520	USA
132	@250HealthyFoods	476	UK
4574	@DialysisSaves	463	USA
6058	@eyerubdesigns	448	USA
11328	@medivizor	371	Global
4445	@diabetesDomin8r	315	USA
7595	@HealthVings	314	INDIA

From the USA, the user named "@DiabeticTalks" is the top significant user. We can observe from the results that most of the active Twitter users for Diabetes are from the USA. Monitoring and following such Twitter accounts will definitely help people to their own health monitoring. From INDIA there is one active member identified i.e "@HealthVings"

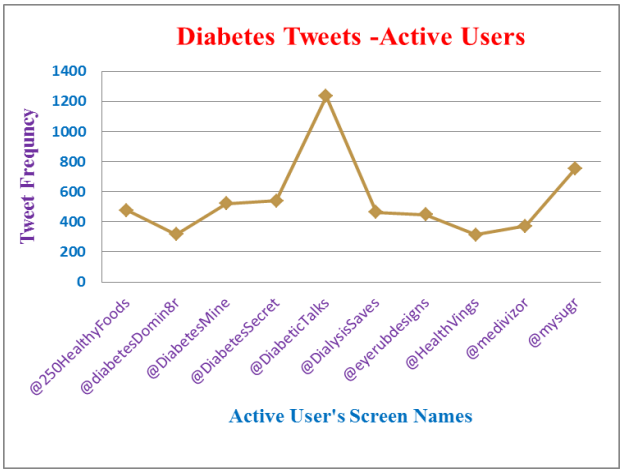


Fig.8.Line plotting Active Users Identified for Diabetes

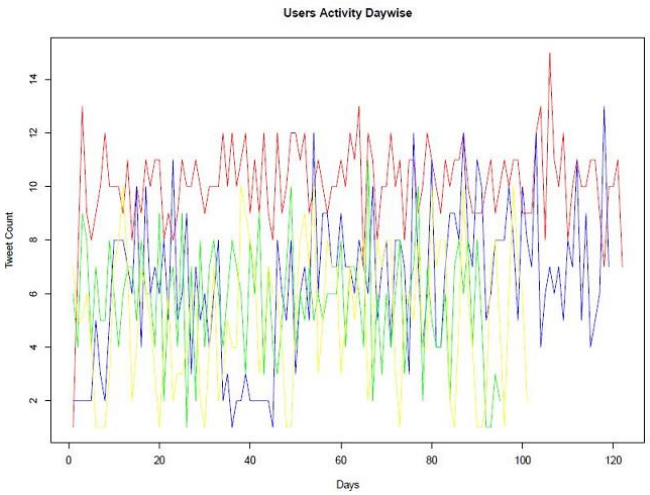


Fig.9. Tweet Activity of Top four Users for a period of three months. X-axis represents dates from Dec2018 to Feb 2019, a period of 120 days. Y-axis represents day wise Tweet Counts for all the Top Users.

C. Exploration Phase tasks

The exploration phase is responsible for two tasks i) Exploring the Popularity of Users ii) Prediction of Growth Rate of Twitter User Activity. This is an extended phase of the previous. Because the derived data from the previous phase is utilized here.

1. Exploring the Popularity

This is the first task of the phase. The task is about the popularity of a Twitter User. The popularity of a Twitter user is defined in terms of no of Twitter followers he/she has. Usually comparing an individual Twitter account with a group account, Group twitter account [10-11] has got more popularity with more followers to it. The extra column is also considered i.e follows count. The classification filtering starts with the tweets file with four fields i.e User's screenName and its tweet frequency count, follower count and follows count. Execution starts with Mapping User's screenName with its followers count and follows count into one Data frame, so that for each screenName Table command is executed in R to find each user wise counts. Bar Chart is used to visualize all the user's Popularity. And is shown in Fig.10.

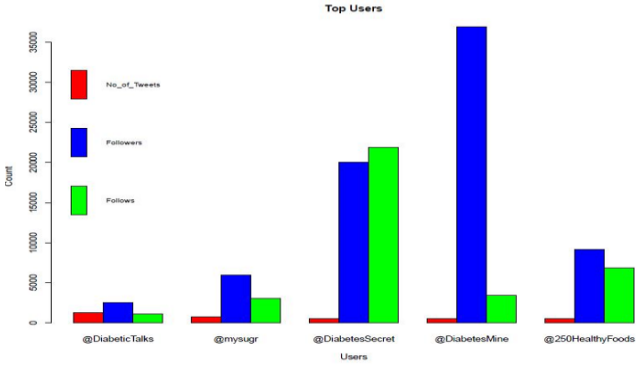


Fig.10. Bar Plot of Diabetes Twitter Users with their Frequency, Followers and Follows counts.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

2. Prediction of Growth rate of Twitter User Activity

In order to predict Popular Twitter Users, the growth rate parameter can be used. By meticulously monitoring the user activity time to time, we can observe some users maintain a regular activity whereas the other is little deviating from the tempo. By considering the cumulative tweet count for each user we can visualize a growth line chart that demonstrates the Top user's activity performance and is shown in Fig.11.

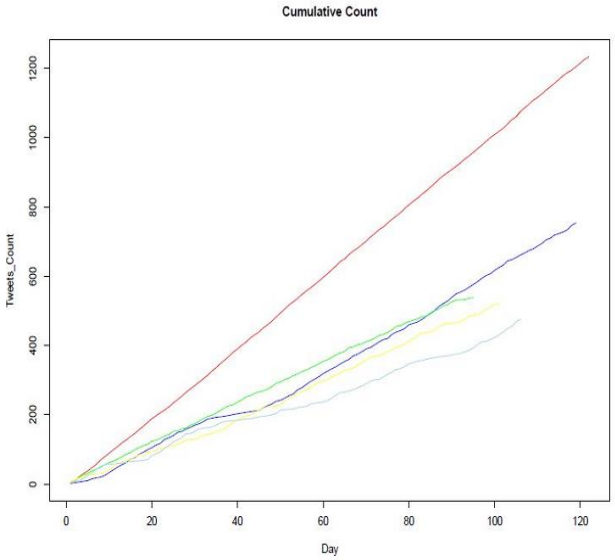


Fig.11. Growth Line chart for the top four Active Users

The growth rate for a period of 30 days for top user1, it is 100%, whereas, for user2, it is 50% and the fourth user it is only 25%. Following the formula 1 in the Exploration phase, the growth rate values are calculated for each user activity. The values are shown in Table 3.

Table 3. No of Tweets posted by different Users every month

	Jan	Feb	March	Growth rate
User1	300	600	900	100%
User2	200	30	450	50%
User3	160	200	300	25%

D. Analysis Phase tasks

This phase includes two subtasks. They are i) Analyzing Positive Tweets ii) Analyzing Negative Tweets. This phase is bound to do sentiment analysis classification of tweets. The Preprocessed Tweets on Diabetes.csv is the input for this phase related tasks. This analysis uses only precise tweets of Diabetes whereas precise tweets are defined as the result of the filtering process on Diabetes tweets having followers count at least 1000. From the filtered tweets file only the tweet text is taken as input for sentiment analysis. The .csv file with only the tweet text is shown in Fig.12.

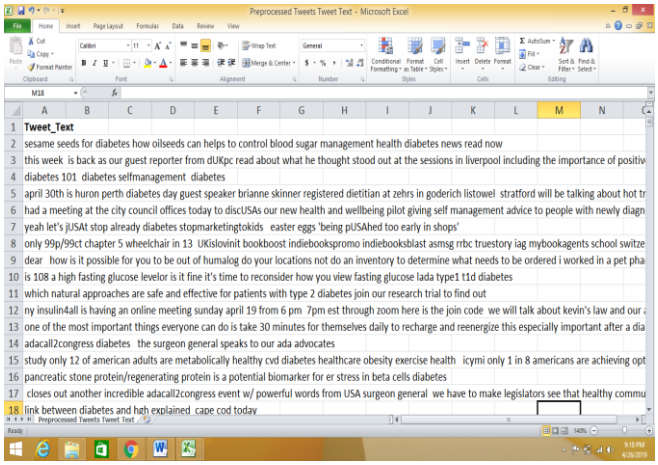


Fig.12. File Input only with Tweet Text

Sentiment Analysis [12] is defined as the process of categorizing tweet into either positive or negative. Regarding Diabetes, all the tweets that write about suggestions regarding food, treatments, etc. classify into the positive category. The Diabetes tweets that talks about causes of Diabetes disease, other affected diseases, and death issues, etc. classify into Negative category. The classification method implemented here for the process of sentiment analysis is Support Vector Machine classification. The SVM Classification [13] starts with classifying the tweets into Training data and Test Data. Training data set of tweets are verified with the predefined lexicons to get a score for each word in the tweet. Then word scores are identified to get a sentiment score for each tweet. Based on the final score, each tweet is labeled as either positive or negative. The sample of the code logic is given below.

model.svm <- train_model(container.svm, "SVM", kernel="linear", cost=1)

The trained tweets data is used to label the test tweets

results.svm <- classify_model(testContainer.svm, model.svm)

The result of the execution of the above code is stored in a .csv file with the fields tweet text, category, score. Negative & positive tweets are stored in different files. These are the input for analysis tasks.

Here the primary aim of this task is to analyze positive tweets of Diabetes to find patterns of suggested foods and negative tweets of Diabetes to find patterns of causes of diabetes and affected diseases because of diabetes.

1. Analysis of Negative Tweets

The negative tweets file is the input for this task. All the negative tweets are endured through tokenization process by executing commands in R. All the tokens are passed through High Ranked frequency Algorithm to find the frequency of each word either a unigram, bigram, trigram. Then filtered to have only words with a frequency threshold. Then manual verification is done to identify the strong patterns of causes identified for Diabetes.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

```
#required libraries
library("stringr")
library("tm")
library("SnowballC")
library("wordcloud")
library("RColorBrewer")
#converting our words into lower case and selecting which set of words is to be used
set_of_words <- tolower(first_set)
#adding that to a dataframe
DF <- as.data.frame(set_of_words)
#finding the occurance of each word in our dataset
for (i in 1:length(set_of_words)) {
DF[i,2] <- sum(str_count(MyData$Tweet_Text,set_of_words[i])) }
```

The resultant causes identified of all three categories unigrams, bigrams, trigrams are put in Table 4. For each word, in the table, its frequency is calculated using the tweet text. The frequency of each word shows its impact level on diabetes. This table data of unigrams, bigrams, trigrams of causes are visualized through a word cloud shown in Fig.13.

Table 4.Identified Patterns of Causes of Diabetes

S.No	Unigrams	Bigrams	Trigrams
1	Hypertension	High Cholesterol	Blood Vessel Damage
2	Obesity	Diet Genes	Parental Fat Diet
3	Stress	Sleep Disturbances	Human Alzheimer's

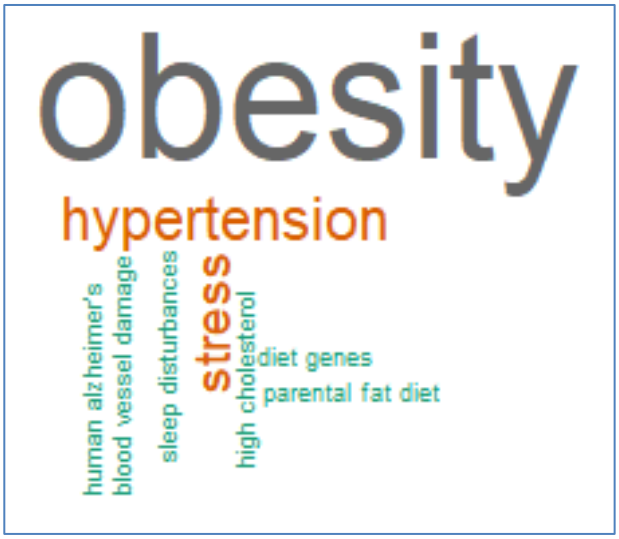


Fig. 13. Word Cloud for Identified Patterns of Causes of Diabetes

The resultant affected diseases identified of all three categories unigrams, bigrams, trigrams are put in Table 5. For each word, in the table, its frequency is calculated using the tweet text. The frequency of each word shows its impact level on diabetes. This table data of unigrams, bigrams, trigrams of causes are visualized through a word cloud shown in Fig.14.

Table 5. Identified Patterns of Affected Diseases of Diabetes

S.No	Unigrams	Bigrams	Trigrams
1	Arthritis	Birth defects	Life threatening issue
2	Infertility	Foot ulcers	Cardio Vascular risk
3	Autism	Heart disease	Asthma respiratory disease
4	Nephrology	Heart stroke	Chemic heart disease
5	Osteoarthritis	Heart Failure	High blood pressure
6		Kidney failure	Health imbalances
7		Organ Transplants	Memory Loss Dementia
8		Muscle weakness	

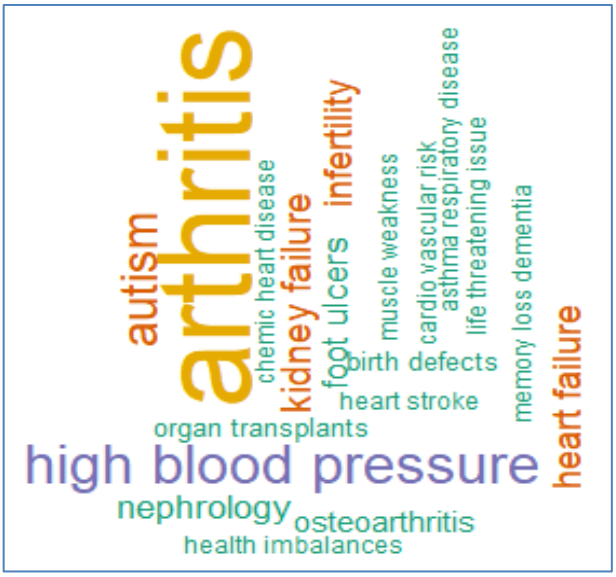


Fig.14. Word Cloud for Diseases Affected by Diabetes

2. Analysis of Positive Tweets

The .csv file with the positive tweets is the input for this analysis. Each tweet is passed through tokenization process to find unigrams, bigrams, trigrams. To get the frequency of all the words, R code is written and executed to find each word with its frequency. Then filtered to get only words meeting the threshold requirement. Here the task is regarding finding food patterns suggested to prevent and control Diabetes. Final filtering is done through once manual verification of all the words. And the final list of unigrams, bigrams, and trigrams are shown in Table 6.

Table 6. Identified Suggested Food Patterns for the Control and Prevention of Diabetes

S.No	Unigrams	Bigrams	Trigrams
1		Herbal remedies	Low carb foods
2	Mushrooms	Personalized diet	Plant based meal
3	Garlic	Salad Vegetable	Plant based nutrition
4	Allicin	Sea weeds	
5	coffee	Turmeric Curcumin	

Preparation of word cloud:

```
#finding the occurance of each word in our dataset
for (i in 1:length(set_of_words)) {
DF[i,2] <- sum(str_count(MyData$Tweet_Text,set_of_words[i]))
}
```


Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

#plotting the wordcloud

wordcloud(words = DF$Words, freq = DF$Freq, min.freq = 0, rot.per=0.4,colors=brewer.pal(8, "Dark2"))

The identified food patterns from table.6. are the input for generating word cloud for Diabetes Food patterns to control. The word cloud is visualized in Fig.15.

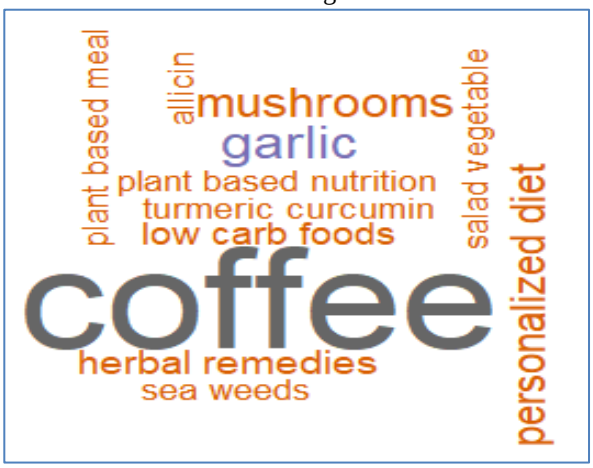


Fig.15. Word Cloud for Suggested Foods for the Prevention and Control of Diabetes

1. CONCLUSION

The framework proposed is suitable for tracking user interests regarding some of the key aspects of diabetes disease using Twitter data. The four modules of work in this paper, demonstrated the retrieval of Diabetes Tweets through Twitter Archiver, Preparing the Tweets ready for experimentation, Identification of Active Locations for Diabetes, Identification of active group users for Diabetes, Finding the popularity of the users, Observing growth rate of top users, Analyzing negative tweets for the identification of causes and affected disease patterns of Diabetes and lastly, analyzing positive tweets for the identification of suggested food patterns of Diabetes. The experiment results from the proposed work proved that the USA stands in the first place in having more internet users and active updation on diabetes, and then UK, INDIA, CANADA, AUSTRALIA, BELGIUM, FRANCE, GERMANY, SINGAPORE in the next order. This paper summarized a detailed overview of Diabetes.

REFERENCES

1 Lauren Sinnenberg, BA, Alison M. Buttenheim, PhD, MBA, Kevin Padrez, MD, Christina Mancheno, BA, Lyle Ungar, PhD, and Raina M. Merchant, MD, MSHP, "Twitter as a Tool for Health Research: A Systematic Review", AJPH RESEARCH, Sinnenberg et al. Peer Reviewed Research e1, January 2017, Vol 107, No. 1 AJPH.

2 Micheal J, Widener , Wenwen Li, "Using geolocated Twitter data to monitor the prevalence of healthy and unhealthy food references across the US" , Applied Geography, 2014 Elsevier.

3 Patrick Park and Michael Macy, "The paradox of active users", Big Data & Society: sagepub.co.uk/journalsPermissions.nav DOI: 10.1177/2053951715606164 bds.sagepub.com, July–December 2015.

4 TariqAhmad, AllanRamsay, HanadyAhmed, "Classification of Tweets using Multiple Thresholds with Self correction and Weighted Conditional Probabilities", New Orleans, Louisiana, June 5–6, 2018. ©2018 Association for Computational Linguistics.

5 Ryan Compton, David Jurgens, David Allen , "Geotagging One Hundred MillionTwitter Accounts with Total Variation Minimization", March 2015.

6 Nathan Eaglea,b,1, Alex (Sandy) Pentlandb, and David Lazerc, "Inferring friendship network structure by using mobile phone data", PNAS September 8, 2009.

7 David A. Broniatowski1,2, Michael J. Paul3*, Mark Dredze4, "National and Local Influenza Surveillance through Twitter: An Analysis of the 2012-2013 Influenza Epidemic", PLOS ONE | www.plosone.org 1 December 2013.

8 Patrick S. Park1,2(•), Ryan F. Compton2, and Tsai-Ching Lu2, " Network-Based Group Account Classification " , © Springer International Publishing Switzerland 2015.

9 Sam Tyner, François Briatte and Heike Hofmann, "Network Visualization with ggplot2", The R Journal Vol. 9/1, June 2017.

10 Jari Saramäkia,1, E. A. Leichtb, Eduardo Lópezb, Sam G. B. Robertsc, Felix Reed-Tsochasb,d, and Robin I. M. Dunbare, "Persistence of social signatures in human communication", PNAS | January 21, 2014.

11 Shaomei Wu, Jake M. Hofman Yahoo, Winter A. Mason , Duncan J. Watts , "Who Says What to Whom on Twitter", ACM, March 28–April 1, 2011, Hyderabad, India.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)