Spring 2015

Online diagnosis of diabetes with Twitter data

Farheen Ali

ONLINE DIAGNOSIS OF DIABETES WITH TWITTER DATA

by

FARHEEN ALI

A THESIS

Presented to the Faculty of the Graduate School of the

MISSOURI UNIVERSITY OF SCIENCE AND TECHNOLOGY

In Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE IN INFORMATION SCIENCE

AND TECHNOLOGY

2015

Approved by

Dr. Fiona Fui-Hoon Nah, Advisor

Dr. Sriram Chellappan, Co-Advisor

Dr. Keng L. Siau

Dr. Michael Gene Hilgers

ABSTRACT

Innovation in technology enables people to communicate, share information and look for their needs by just sitting in rooms and going through some clicks. While social media has played a very important role in connecting people worldwide, its potential has stretched beyond the innovative idea of connecting people through their social networks. While many thought there was no meeting point for the healthcare sector and social media, it was a surprise when research and innovations have shown that social media could lay a very significant role in the health care sector.

Research has been done in developing models that could use social media as the data source for tracking diseases. Most of these analyses are based on models that prioritize strong correlations with seasonal and pandemic kinds of diseases over the health conditions of a specific individual user.

The aim of this research is to develop a diabetes detecting tool at the individual level using a sample of Twitter IDs that have been collected from the Twitter search using the query –'*recently diagnosed'* and *'diabetes'*. Based on text analysis of social media posts using Fisher's exact test, without any medical settings, this thesis investigates the feasibility of diagnosing and classifying diabetes via machine learning techniques, Naive Bayes and Random Forest classifiers. It was found that more than half (20/30 ≈ 67%) of the users in the sample mentioned being tested positive for diabetes, about 27% (8/30) of the users mentioned the symptoms and got involved in diabetes related discussions, but did not mention about being tested positive and rest 4% had no mention of symptoms or diabetes.

ACKNOWLEDGMENT

My first thanks and heartfelt gratitude goes to Dr. Fiona Fui-Hoon Nah, my advisor, for giving me the freedom to pursue my own interests and for trusting me on the same. I could not have completed this thesis without her valuable suggestions and those brainstorming meetings, where she taught me how to assess a problem and find a best possible solution to it. I would like to thank her for being so patient with me, and helping and guiding me to improve this thesis and in bringing it to this shape.

I would also like to thank my co-advisor, Dr. Sriram Chellappan, for introducing me to the concept of *Health Diagnosis via Social Network*, offering me his invaluable assistance despite his busy schedule, and for discussing with me his innovative ideas. Without his motivation and support, I wouldn’t have been able to learn about this topic and get a deeper understanding. I would also like to thank Dr. Keng L. Siau and Dr. Michael Gene Hilgers for being part of my thesis committee and taking time to review this work. This thesis would not be possible without the generous help of Raja Ashok Bolla, who helped me by providing the tweets from the filtered Twitter IDs.

I saved the last for people closest to my heart – my family. I’m very thankful to my parents, Dr. Mir Firman Ali and Shahnavaj Begum, and my siblings, Dr. Syed Irfan Ali and Dr. Nasreen Ali, for helping me understand the medical terms and concepts related to diabetes and for being patient with me while I dragged I.T. to medical science and questioned a few traditional concepts. I specially wish to acknowledge Dr. Sekh Ansar Alli, my brother-in-law, for encouraging me to pursue a master’s degree. If it weren’t for him, I would have missed out on this amazing experience.

TABLE OF CONTENTS

Page

SECTIONS

APPENDICES

LIST OF ILLUSTRATIONS

Figure Page

LIST OF TABLES

Table Page

1. INTRODUCTION

This section begins by stating the problem description and motivation for conducting this research. This is followed by the main research question and a very brief outline of the proposed research approach. The section closes with an outline of this thesis along with the major research contributions.

1.1 PROBLEM DESCRIPTION

A human body consumes energy to perform different daily tasks. The source of this energy is the food that is consumed. An organ called the pancreas, in a human body, lying near the stomach, produces a hormone called insulin, which helps glucose to reach all the cells of a human body. Diabetes is a metabolic disease, in which either the body fails to make sufficient insulin or cannot utilize the insulin the way it should, which in return causes sugar to build up in the body. Diabetes, if not controlled, causes complications and effects heart, nerves, eyes, feet and kidneys [1].

The early common symptoms of diabetes include [2]:

- Frequent urination
- Feeling very thirsty
- Frequently feeling hungry
- Extreme fatigue
- Blurry vision
- Cuts/bruises that are slow to heal
- Weight loss - even though a person eats more (type 1)
- Tingling, pain, or numbness in the hands/feet (type 2)

According to statistics, approximately 366 million people suffer from diabetes and it is estimated that by the end of 2030, the count will rise to 552 million. It is a known fact that an early diagnosis of diabetes can help prevent progression to later complications. Research states that about 183 million people presently have diabetes and are unaware of it. One of the types of diabetes, type 2, can be evident in people for about 9 – 12 years without their knowledge and can cause complications during treatment. **Early detection of diabetes is crucial for active management for people who have been newly diagnosed and have not developed complications yet** [3]. It is unlikely to expect everybody to be aware of the early symptoms of diabetes and visit a doctor. However, in today’s world, according to “Worldwide Social Network Users: 2013 Forecast and Comparative Estimates”, approximately one in four people across the globe use social networks, and this number is believed to have risen from 1.47 billion in 2012 to 1.73 billion in 2014, with an estimated 18% increase [4].

Twitter is one of the most famous online social networking services, with an estimated 310,000,000 monthly visitors and 500 million users worldwide [5]. Within a character limit of 140, it allows its users to post their thoughts and opinions, and gives its registered users the privilege to read and comment [6]. Twitter provides its users a platform to converse on almost every topic known to man, and thus, people started discussing their health intentionally or unintentionally as well. This has intrigued many researchers to look into the most common diseases people discuss and the scope of the possible diagnosis virtually.

This study, hence, focuses on a potential system which can help a healthcare professional to track his/her patients' Twitter posts and diagnose diabetes in accordance with the symptoms they post, with the help of social network analysis and text analysis.

1.2 SOCIAL MEDIA AND HEALTHCARE: AN OVERVIEW

Social media is a group of internet-based applications developed using Web 2.0 technology that offer opportunities for users to generate, share, receive, and comment on social content among multiuser through multisensory communication [7, 8, 9, 10, 11]. Research has shown that there is a relationship between personality traits and engagement with social media [12].

Social media brings a novel dimension to health care, as it proposes a medium to be employed by the public, patients, and health professionals to communicate about health issues with the possibility of potentially improving health outcomes. Social media is changing the nature and speed of health care interaction between individuals and health organizations. The general public, patients, and health professionals are using social media to communicate about health issues [13] including health promotion and health education [14, 15, 16, 17]. Social media has widened access and increased awareness among those who may not easily access health information via traditional methods, such as younger people, ethnic minorities, and lower socioeconomic groups [18, 19, 20]. Colineau and Paris [21] from their research have reported that people prefer using health-related social networking sites to discuss sensitive issues and complex information with health professionals. One of the advantages of social media is that it lets the health information reach its audience via various other modes than just text; for example, videos can be used to replace text and can be useful when literacy is low [22].The very famous video sharing website YouTube allows users to share, upload and view videos online for free, has been used by the general public to share and learn about medications, symptoms, and diagnoses [21] and by patients to share personal cancer stories [23].

Social media adoption rates have shown variations in accordance to the geographic locations; for example, in Europe the percentage of German hospitals using social networks are in “single figures”, whereas approximately 45% of Norwegian and Swedish hospitals are using LinkedIn, and 22% of Norwegian hospitals use Facebook for health communication [24]. In the United States, on the other hand, 61% of adult search online and 39% use social media such as Facebook for health information [25]. The growth in popularity of social media among the general public has caused the research and evolution of many applications within health contexts, ranging from the World Health Organization using Twitter during the influenza A (H1N1) pandemic, with more than 11,700 followers [26], to medical practices [27], and health professionals obtaining information to inform their clinical practice [28, 29].

There is a range of social media platforms available currently that can facilitate a dialogue between patients and health professionals [21, 30]. For example, sites such as PatientsLikeMe enable patients to easily converse with others and share health information and advice including information on treatment and medication [31, 32]. Famous social networking sites such as Twitter and Facebook are being used by the general public, patients, and health professionals to share their experience of disease management, exploration, and diagnosis [33]. Blog sites create a space where people can access tailored resources [34] and provide health professionals with an opportunity to share information with patients and members of the public [35, 21]. Asthma groups are using MySpace to share health information, in particular personal stories and experiences [36, 37].

Nowadays, social media is been used by many researchers to collect data on patient experiences and opinions such as symptoms, physician's performance etc. [34, 38]. With the help of these new modes of interactions, social media can monitor public response to health issues [20], track and monitor disease outbreaks [39], identify target areas for intervention efforts [40], and disseminate pertinent health information to targeted communities [41]. Health professionals can aggregate data about patient experiences from blogs and monitor the public reaction to health issues.

1.3 RESEARCH QUESTION AND MAJOR CONTRIBUTIONS

In the light of the reasons described in section 1, the main research question addressed by this thesis is as follows:

Is it possible to observe diabetes based on text analysis of social media even if the individual does not intentionally discuss his/her health?

While there is a lot of existing work on prediction of seasonal and pandemic diseases, to the best of the author's knowledge, an attempt to diagnose a non-seasonal as well as a non-pandemic disease, like diabetes, based on an individual's post on Twitter has never been done before. Considering the fact that non-pandemic diseases are extremely complex due to their similarity of early symptoms with many other diseases, it becomes very difficult to possibly predict one without having actually met the patient in person [42]. This thesis, therefore, presents an original work with the following as its major contributions:

- **Approach to deal with the problem statement:** One of the previous works done on healthcare and social media includes the prediction of Influenza and Influenza-like (ILI) activity in the USA, prior to the generation of Control and Prevention (CDC) report and the source of the data has been Wikipedia usage and individual based tweets.

  Both of the above approaches demonstrate the diagnosis of seasonal and pandemic diseases where the trends and estimation of the time period plays an important role in the determination [43]. This research, on the other hand, focuses on diseases which are not seasonal, yet the individuals who have been diagnosed with these diseases have shown similar trends of mentioning the symptoms in their tweets.

- **Approach to solve the problem statement:** Twitter provides its API (Application Programming Interface) to researchers and other web developers, and hence allows a web platform to access and share information from one another.

  My approach towards solving this problem involves, with the help of these Twitter APIs, collecting all the individual posts, filtered from the IDs based on the Twitter search query which involves keywords '*recently diagnosed*' and '*diabetes*' and then sorting them as cases and then comparing them to a sample of randomly selected subset of people without the attribute (the controls) [44]. The entire

Twitter history of these individuals was extracted, and a model was developed to count the number of times they posted the symptoms (such as sleep, water, eye, rash, tired, etc.) related to diabetes [45], which being the early symptoms of diabetes, the trends for these symptoms match the curve obtained from searching the keyword ‘diabetes’ in Google trends. The more the number of keywords mentioned in the posts, over a period of time, the greater the probability of a person being possibly diabetic.

- **Diabetes, if known in the early stage, can help to take prior precautions and keep the blood sugar level in control:** A solution to this problem statement is important since it would provide a patient more time, if detected earlier, to control diabetes and prevent it from getting worse.

1.4 THESIS ORGANIZATION

Excluding this section, the remainder of the thesis is organized in the form of four sections.

Section 2 describes the research methodology along with the necessary processes and infrastructure used to obtain the outcome from the hypothesis. This section also describes the research approach used, by dividing the entire thesis into three phases and summarizing each of them.

Section 3 describes the process used to collect, clean and parse the relevant data. This section closes by presenting a detailed account of various statistical insights obtained by performing statistical analysis on the extracted data.

Section 4 is solely devoted to the main research task of building classification classifiers. This section begins with sorting the training data for Naive Bayes classifier and then creating the probability table. The later portions of this section describes how to build a Random Forest method of classification. The section concludes by presenting performance metrics such as the accuracy and precision for the proposed classification models.

Finally, Section 5 concludes the thesis by summarizing all the four sections and the results obtained. In addition, several potential approaches are described for improving the classification accuracy, intended as a guide to future researchers who wish to extend and build on this thesis.

2. RESEARCH METHODOLOGY

This section begins with the description of the types of analysis performed in this study along with the detailed description of the necessary research infrastructure. This is followed by a brief description of the proposed research methodology.

2.1 FISHER'S EXACT TEST

Named after its inventor, Sir R. A. Fisher, Fisher's exact test is a statistical significance test used for a 2 X 2 contingency table in order to compare the binomial probabilities and to test for independence of 2 classifications. It is believed that Fisher's exact test helps to exactly calculate the deviations from a null hypothesis, independent of the sample size or the sample characteristics, hence it falls under the class of exact test. Although it is valid for all sample sizes, because of the above grounds, this test is preferred when sample sizes are small [46]. The purpose of using the Fisher's exact test is to classify the categorical data in order to determine the significance of contingency between them. The null hypothesis for Fisher's exact test states that assuming each observation is classified into exactly one cell and the rows and columns are fixed, the comparative proportions of two variables are independent of each other. In simpler terms, there is no affiliation between the rows and columns of a 2 X 2 contingency table, such that the probability of a subject being in a particular row is not determined by being in a particular column [47]. Provided the margins are fixed, the Fisher's exact test when applied to a table with cells a, b, c & d and the marginal totals (a + b), (c + d), (a + c) and (b + d):

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!\ (c+d)!\ (a+c)!\ (b+d)!}{a!\ \ b!\ \ c!\ \ d!\ \ n!}$$

Where $\binom{n}{k}$ is the binomial coefficient.

2.2 NAIVE BAYES CLASSIFIER

Descending from the family of simple probabilistic classifiers, Naive Bayes is a popular method for text classification i.e. it judges the belonging of documents in their respective categories (such as sports or politics, healthy or sick etc.) on the basis of word frequencies as the features [48]. Based on the Bayesian theorem, this classifier assumes the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. For example, an orange is a fruit with distinctive features of orange in color, round and about 4’ in diameter. Now irrespective of other features present or the fact that these features may be dependent on each other, a Naive Bayes classifier would consider all of these properties to independently contribute to the probability that the given fruit is an orange. This type of classifier is henceforth useful in medical diagnosis, since it would work very well with diseases showing similar symptoms. Further, it is also capable of working well with a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. Other advantages of using Naive Bayes classifier include its non-sensitivity to irrelevant features, its capability to handle real, discrete and streaming data and most importantly it is fast to train and classify. Naive Bayes classifier is particularly suited when the dimensionality of the input is higher. Parameter estimation of this model uses the method of maximum likelihood [49].

Likelihood

Class Prior Probability

$$P(c \mid x) = \frac{P(x \mid c)\,P(c)}{P(x)}$$

Posterior Probability

Predictor Prior Probability

$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

- P(c|x) is the posterior probability of class (target) given predictor (attribute).
- P(c) is the prior probability of class.
- P(x|c) is the likelihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor. [14]

2.3 RANDOM FOREST

Random forest is an ensemble learning method developed by Leo Breiman and Adele Cutler [50], which during its training period constructs a magnitude of decision trees and outputs the resultant mode of the classes (classification) or mean prediction (regression) of the individual trees. Dietterich first came up with the idea of randomized node optimization, where instead of deterministic optimization, a randomized procedure was used to select decisions from each node. Usually used for classification and regression, this method is better than the decision trees since they do not over fit their training data by providing too many parameters relative to the number of observations [50]. Unlike the standard tree methods of classification, in the case of random forest method, the best among a subset of predictions are chosen randomly to split each node [51]. This method is considered to be more user-friendly since it has only two parameters (the number of variables in the random subset at each node and the number of trees in the forest), and is usually not very sensitive to their values. The concept behind this method is growing a forest of trees and inducing the variation among the trees by projecting the training data into a randomly chosen subspace before fitting with each other. It averages multiple deep decision trees during this process with an intention to reduce the variance and during this it also boosts the performance of the final model [52]. An estimate of the error rate can be obtained based on the training data by calculating the out-of-bag error.

Out-of-bag error

While in the process of retaining the training set by sampling with replacement for the current tree, about one-third of the cases are left out of the sample and not used in the construction of the kth tree. This left out data is termed as out-of-bag data and is utilized to get an unbiased estimation of the classification error as the trees are added to the forest. Once the tree is built, the entire set of data is made to go through the tree and in the meanwhile the proximities for each pair cases are computed. The proximity increases by one if any of the two cases occupy the same terminal node. The proximities are then normalized in the end by dividing with the number of trees. Now, let us consider j is a class that scored more points every time case n was out of the bag. Thus, the proportion of the time that j doesn’t match with the class of n averaged over all cases is the out-of-bag error estimate [51, 52].

2.4 RESEARCH APPROACH

This thesis is accomplished by a three phase procedure.

- **Phase I - Data collection from social network analysis:** A Twitter search query, including keywords, “diagnosed, diabetes” is used to search for all the Twitter IDs of people who posted about being diagnosed with diabetes. These IDs were then filtered to find out only those prospects who posted about being diagnosed recently. The resultant Twitter IDs were made to pass through a Java program, which would process posts of each ID, one at a time and count the number of times the symptoms of diabetes have been mentioned in the past.
- **Phase II - Data preprocessing and statistical analysis:** Fisher’s exact test was used to find the similarity between the symptoms, used as the keywords for selecting the prospective Twitter profiles, and diabetes. The Twitter profiles were narrowed down on the basis of these keywords used in the posts along with a mention of being diagnosed with diabetes.
- **Phase III - Diabetes classification using Machine Learning Techniques:** Leveraging on the statistical insights from phase II, two machine learning techniques (Random Forest classification method & Naive Bayes’ classifier) will be employed to perform diabetes classification.

3. TWITTER DATA PROCESSING

This section provides a detailed explanation of the data collection from Twitter, data processing and statistical analysis over the collected data. The analysis of the posts on Twitter used to determine the symptoms of diabetes consists of the following steps:

1. Collection of tweets.
2. Cleaning and parsing of data.
3. Conducting statistical analysis of the extracted data.

3.1 COLLECTION OF TWEETS

Twitter welcomes developers to explore its platform for research purposes. An unofficial Java library for the Twitter Application Program Interface (Twitter API), provides application automation so that it can be integrated with Twitter. A few healthcare professionals were contacted and consulted in order to have a better understanding about diabetes and its early symptoms. Tweets were collected on the basis of the symptoms suggested by the physicians. A keyword strategy was adopted for the collection purposes. The following symptoms were derived from the discussions: sleeping disorder, obesity, water loss in the body, susceptibility to heat, and the redundant need of eating food. From these early symptoms, the following keywords were then derived: sleep, weight, water, heat, hungry; and these keywords were looked up in the Twitter search to find the IDs of people who have posted the same keywords in the past along with the phrases ‘recently diagnosed’ and ‘diabetes’. Only those prospects who have mentioned at least 2 or more symptoms in their Twitter posts (for accuracy purpose, mentions of minimum two symptoms were looked up) along with the mention of being recently diagnosed with diabetes, were taken into account and then passed though the Java code to count the number of times these symptoms were posted and the dates of their postings. A sample of tweets collected, where users have unintentionally tweeted their symptoms and with time, eventually mentioned about diagnosed diabetic, are shown in Table 3.1.

Table 3.1. Sample Tweets Collected

S. No.	**Date**	**Tweet**
1.	Mon Apr 14 12:12:57 CDT 2014	Hurting, need more *sleep*
	Mon Jul 21 01:58:04 CDT 2014	Finally ate something today, but I did drink a lot of *water*, as always
	Wed Jul 30 21:55:29 CDT 2014	I wilt in this *heat*
	Sun Jul 20 17:55:37 CDT 2014	I am so *hungry*, having a :(kinda day.
	Fri Sep 12 00:57:28 CDT 2014	Kinda like my name:) This *diabetes* sight has some great quotes. Recently found it, and was *recently diagnosed*.

Table 3.1. Sample Tweets Collected (cont.)

2.	Sun Mar 03 17:45:55 CST 2013	*Sleep*, who needs it sef. Back in the groove
	Wed Jun 05 23:15:29 CDT2013	My *weight* loss journey started on Monday
	Fri Aug 23 09:52:32 CDT 2013	Why y run? Couldn't stand the *heat*
	Wed Aug 13 00:06:35 CDT2014	*Recently diagnosed* with *diabetes*, this story scares me
3.	Sat Aug 02 02:24:30 CDT 2014	Okay for real, someone please pray/will/voodoo me to *sleep*.
	Wed Sep 24 23:09:42 CDT 2014	So either I freeze to death tonight or I die of *heat* exhaustion. Either way, I will not live to see 8 am.
	Mon Sep 15 11:18:20 CDT 2014	Love feeling so *hungry* and nauseous at the same time.
	Fri Oct 17 19:18:36 CDT 2014	I was *recently diagnosed* with *diabetes* so I am trying to be good.

The results from the execution of the search query included both relevant and irrelevant accounts (such as the accounts by diabetes community, diabetes association and some by the nurses or diabetes physicians). The accounts which were irrelevant were discarded and only the potential accounts were taken into consideration. Out of 30 potential Twitter IDs obtained, 20 had claimed to have been recently diagnosed with diabetes, and the rest 10, even though they mentioned the symptoms but were not diabetic or claimed to have diagnosed. These 30 Twitter IDs were hence taken into consideration and further analysis was performed on them.

3.2 CLEANING AND PARSING DATA

All the tweets obtained were parsed before Fisher's exact test was conducted. The procedure for parsing included the following steps:

1. Individual terms in a tweet were separated according to the white space boundaries.
2. The tweets were then converted into lower case letters.
3. Finally, all the non-alpha numeric characters were removed from tweets (e.g., hash signs and dashes).

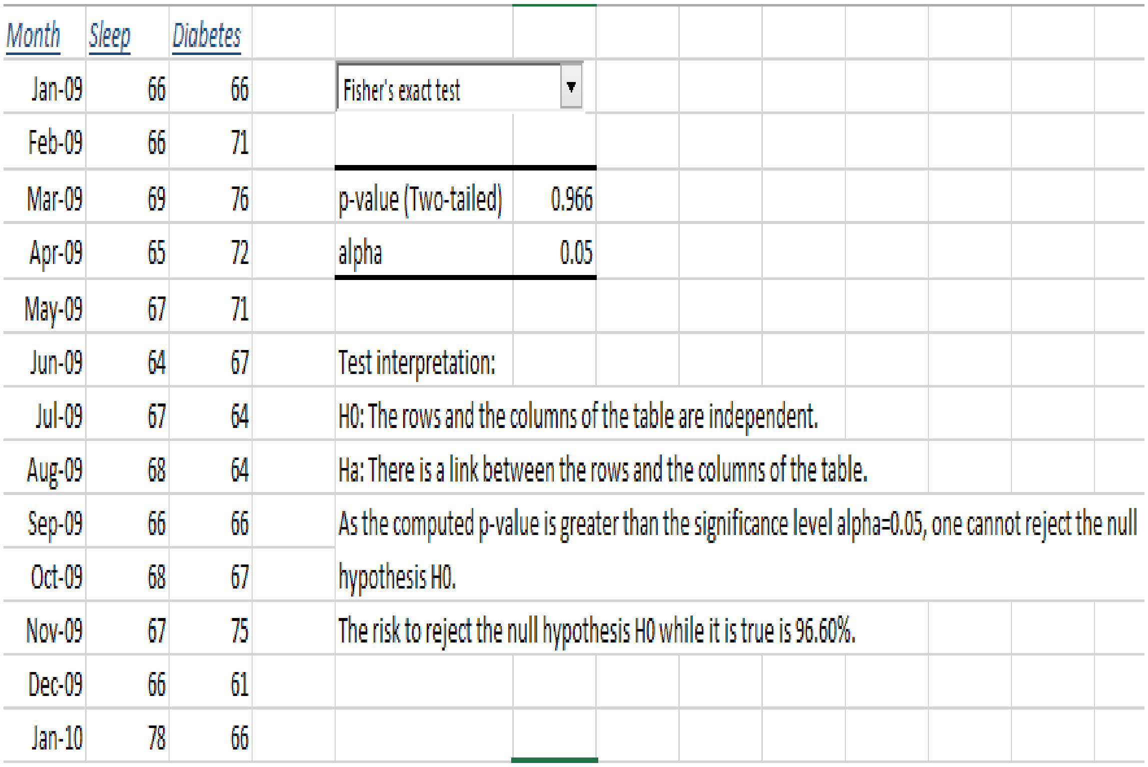
After parsing the tweets, a count function was used to count the number of times the keywords were used by each user in his/her posts. This helped to sort the Twitter IDs confidently according to their relevancy.

3.3 CONDUCTING STATISTICAL ANALYSIS

Fisher's exact test was used to analyze the statistical significance of the contingency table with a primary goal to find out the trends. Fisher's exact test is applied to the derived keywords to assess if each keyword and ‘diabetes’ exhibit similar trends. The results showed that the keywords sleep, as shown in Figure 3.1, water, as shown in Figure 3.2, rash, as shown in Figure 3.3, and tired, as shown in Figure 3.4, show similar trends as diabetes. Other keywords like heat, hungry and itch, do not show similar trends as diabetes, but they are commonly found in the Twitter posts of those who possibly have diabetes.

The data for diabetes and the corresponding keywords are collected for the years 2009 – 2013 and because of the space constraints, limited data can be viewed in the following screenshots.

The results are as follows:



Month	Sleep	Diabetes
Jan-09	66	66
Feb-09	66	71
Mar-09	69	76
Apr-09	65	72
May-09	67	71
Jun-09	64	67
Jul-09	67	64
Aug-09	68	64
Sep-09	66	66
Oct-09	68	67
Nov-09	67	75
Dec-09	66	61
Jan-10	78	66

Fisher's exact test

p-value (Two-tailed)	0.966
alpha	0.05

Test interpretation:

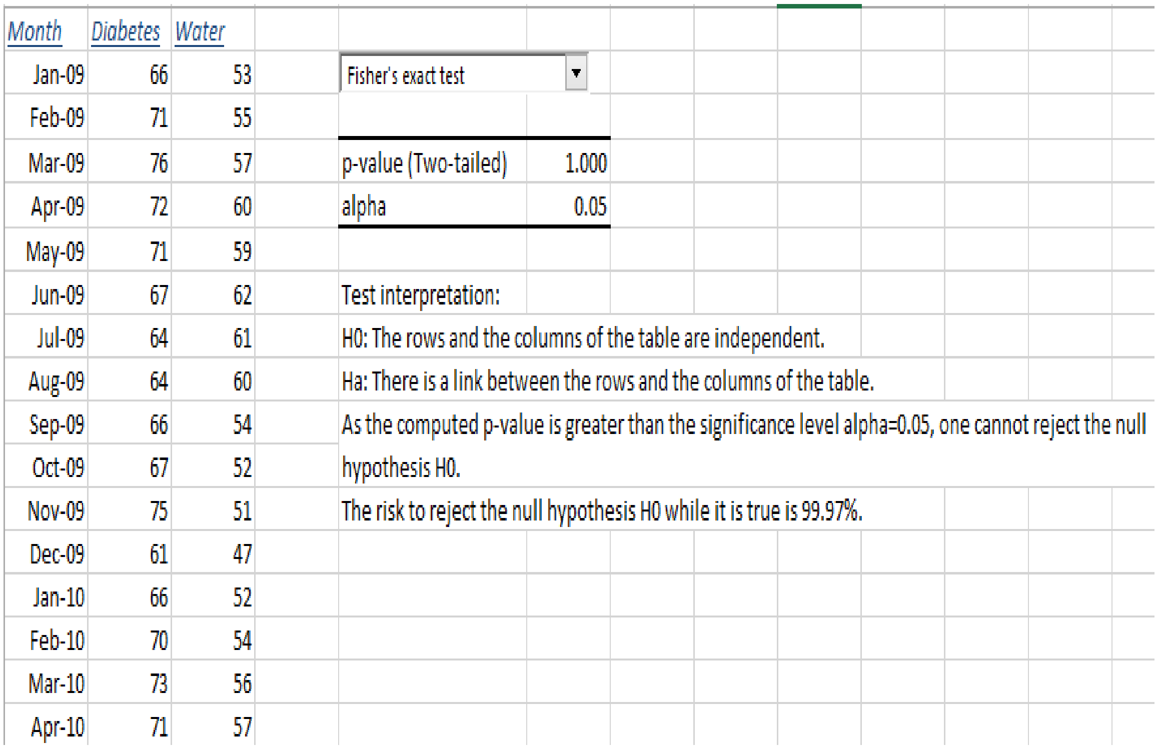
H0: The rows and the columns of the table are independent.

Ha: There is a link between the rows and the columns of the table.

As the computed p-value is greater than the significance level alpha=0.05, one cannot reject the null hypothesis H0.

The risk to reject the null hypothesis H0 while it is true is 96.60%.

Figure. 3.1. Fisher’s Exact Test For Diabetes & Sleep



Month	Diabetes	Water
Jan-09	66	53
Feb-09	71	55
Mar-09	76	57
Apr-09	72	60
May-09	71	59
Jun-09	67	62
Jul-09	64	61
Aug-09	64	60
Sep-09	66	54
Oct-09	67	52
Nov-09	75	51
Dec-09	61	47
Jan-10	66	52
Feb-10	70	54
Mar-10	73	56
Apr-10	71	57

p-value (Two-tailed)	1.000
alpha	0.05

Test interpretation:

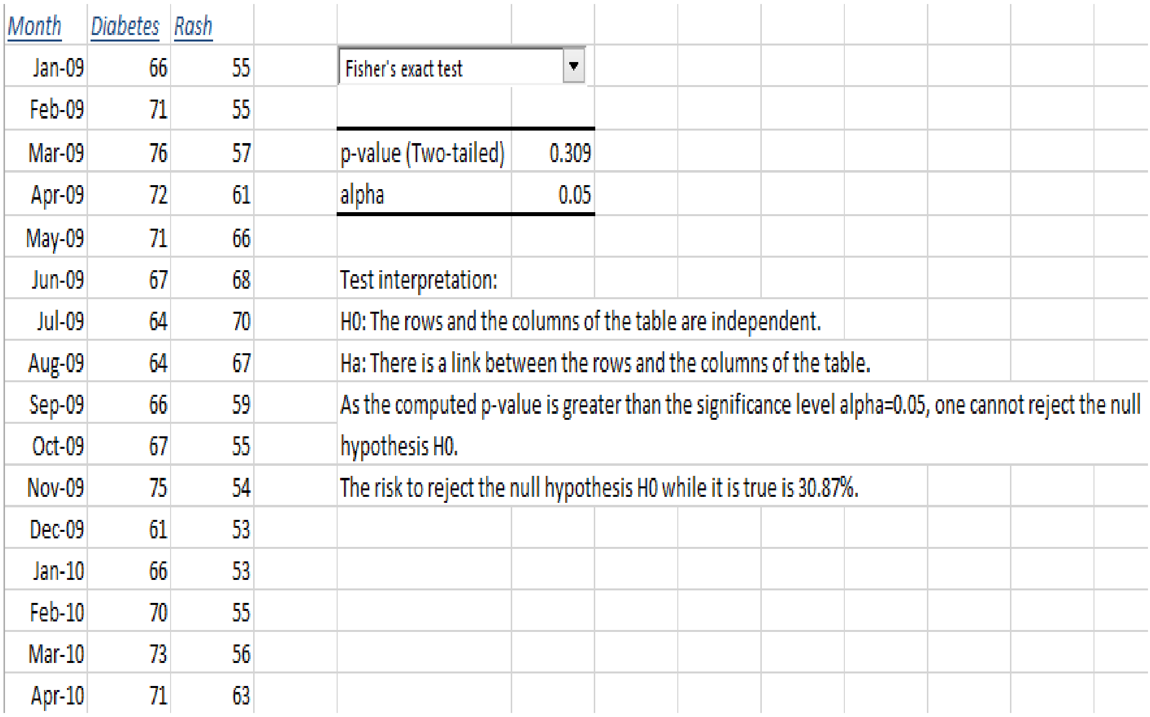
H0: The rows and the columns of the table are independent.

Ha: There is a link between the rows and the columns of the table.

As the computed p-value is greater than the significance level alpha=0.05, one cannot reject the null hypothesis H0.

The risk to reject the null hypothesis H0 while it is true is 99.97%.

Figure. 3.2. Fisher's Exact Test For Diabetes & Water



Month	Diabetes	Rash
Jan-09	66	55
Feb-09	71	55
Mar-09	76	57
Apr-09	72	61
May-09	71	66
Jun-09	67	68
Jul-09	64	70
Aug-09	64	67
Sep-09	66	59
Oct-09	67	55
Nov-09	75	54
Dec-09	61	53
Jan-10	66	53
Feb-10	70	55
Mar-10	73	56
Apr-10	71	63

p-value (Two-tailed)	0.309
alpha	0.05

Test interpretation:

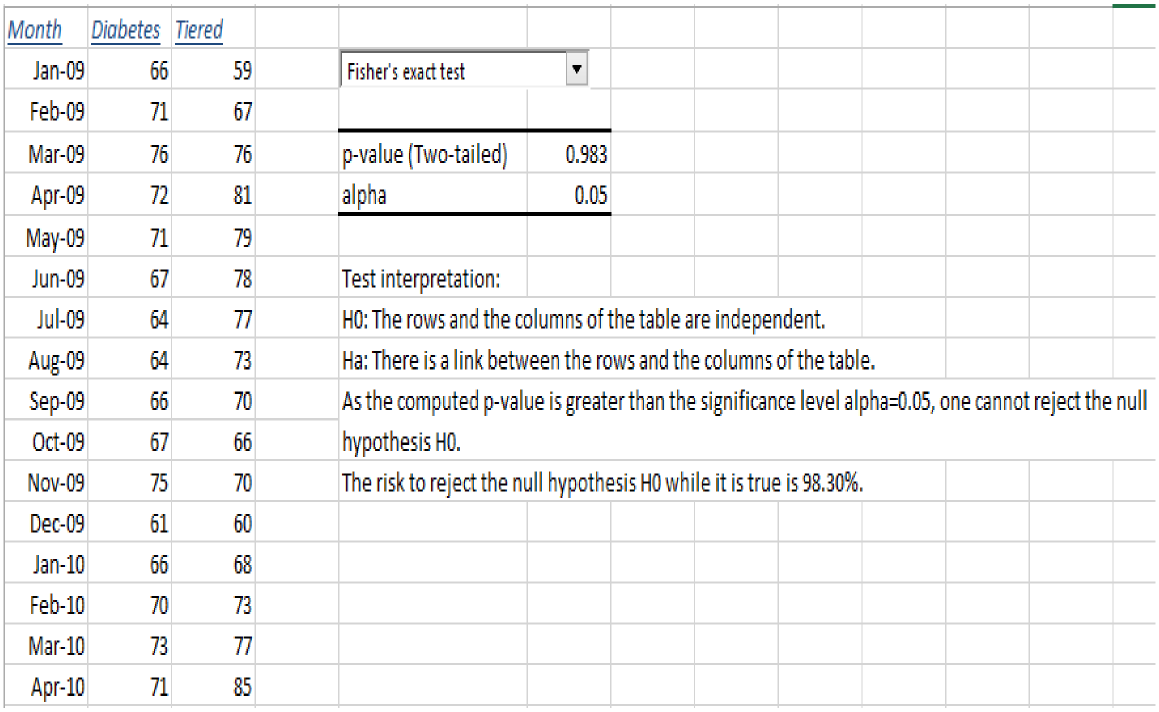
H0: The rows and the columns of the table are independent.

Ha: There is a link between the rows and the columns of the table.

As the computed p-value is greater than the significance level alpha=0.05, one cannot reject the null hypothesis H0.

The risk to reject the null hypothesis H0 while it is true is 30.87%.

Figure. 3.3. Fisher's Exact Test For Diabetes & Rash



Month	Diabetes	Tiered
Jan-09	66	59
Feb-09	71	67
Mar-09	76	76
Apr-09	72	81
May-09	71	79
Jun-09	67	78
Jul-09	64	77
Aug-09	64	73
Sep-09	66	70
Oct-09	67	66
Nov-09	75	70
Dec-09	61	60
Jan-10	66	68
Feb-10	70	73
Mar-10	73	77
Apr-10	71	85

Fisher's exact test ▼

p-value (Two-tailed)	0.983
alpha	0.05

Test interpretation:

H0: The rows and the columns of the table are independent.

Ha: There is a link between the rows and the columns of the table.

As the computed p-value is greater than the significance level alpha=0.05, one cannot reject the null hypothesis H0.

The risk to reject the null hypothesis H0 while it is true is 98.30%.

Figure. 3.4. Fisher's Exact Test For Diabetes & Tired

In all the above cases, one cannot reject the null hypothesis, and therefore, it can be concluded that these keywords show a similar trend as 'diabetes'.

4. MACHINE LEARNING TECHNIQUE AND RESULTS

In this section, an overview of the supervised machine learning technique used for classification is described. The parsed data is analyzed to classify a user as being sick or not.

4.1 NAIVE BAYES CLASSIFIER

Naive Bayes classifier assumes that all the features are independent of each other i.e. the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. Henceforth, this classifying technique is used to determine on the basis of the training data, if it can predict diabetes in the testing data confidently. Naive Bayes classifier formula is applied as discussed in Section 2.2 to the training data, as shown in Table 4.1, and perform the computations as in Table 4.2.

Table 4.1. Training Data For Naive Bayes Classifier

Sleep	**Water**	**Weight**	**Heat**	**Hungry**	**Diabetes**
Yes	No	No	Yes	No	Yes
Yes	Yes	No	Yes	Yes	Yes
Yes	No	No	Yes	Yes	Yes
Yes	No	Yes	No	No	Yes
Yes	Yes	Yes	Yes	Yes	Yes
Yes	Yes	No	Yes	Yes	Yes
Yes	No	No	Yes	No	Yes

Table 4.1. Training Data For Naive Bayes Classifier (cont.)

Yes	Yes	Yes	Yes	Yes	Yes
No	Yes	No	Yes	No	Yes
Yes	Yes	Yes	Yes	Yes	Yes
Yes	No	Yes	No	No	Yes
Yes	Yes	No	No	No	Yes
Yes	Yes	Yes	Yes	No	Yes
Yes	Yes	Yes	Yes	Yes	Yes
Yes	Yes	Yes	Yes	No	Yes
Yes	Yes	Yes	Yes	Yes	Yes
Yes	Yes	Yes	Yes	Yes	Yes
Yes	Yes	Yes	Yes	Yes	Yes
Yes	Yes	Yes	Yes	Yes	Yes
Yes	Yes	No	Yes	Yes	Yes
Yes	No	No	Yes	No	No
Yes	Yes	No	Yes	No	No
No	Yes	Yes	Yes	No	No
No	No	No	No	No	No
Yes	Yes	No	Yes	Yes	No
Yes	Yes	Yes	Yes	Yes	No
Yes	Yes	Yes	Yes	Yes	No
Yes	Yes	Yes	Yes	Yes	No
Yes	No	No	No	No	No
No	No	No	No	No	No

Table 4.2. Probability Table From Training Data

P (Diabetes=yes) = 0.66	P (Diabetes=no) = 0.33
P (Sleep=yes | Diabetes=yes) = 0.95	P (Sleep=yes | Diabetes=no) = 0.7
P (Sleep=no | Diabetes=yes) = 005	P (Sleep=no | Diabetes=no) = 0.33
P (water=yes | Diabetes=yes) = 0.7	P (water=yes | Diabetes=no) = 0.6
P (water=no | Diabetes=yes) = 0.25	P (water=no | Diabetes=no) = 0.4
P (weight=yes | Diabetes=yes) = 0.6	P (weight=yes | Diabetes=no) = 0.4
P (weight=no | Diabetes=yes) = 0.4	P (weight=no | Diabetes=no) = 0.6
P (Heat=yes | Diabetes=yes) = 0.85	P (Heat=yes | Diabetes=no) = 0.7
P (Heat=no | Diabetes=yes) = 0.15	P (Heat=no | Diabetes=no) = 0.3
P (Hungry=yes | Diabetes=yes) = 0.6	P (Hungry=yes | Diabetes=no) = 0.4
P (Hungry=no | Diabetes=yes) = 0.4	P (Hungry=no | Diabetes=no) = 0.6

The sample/testing data is as follows:

Weight = No, Sleep = Yes, Water = Yes, Heat = Yes, Hungry = Yes

It is known that this sample mentioned about testing positive for diabetes in his/her posts, but in order to check if the classifier works as expected, the above probability values for each keyword/symptom obtained from the training data, is used to verify the probability of diabetes in the new sample.

Thus, according to the Naive Bayes' classifier formula:

P (Diabetes=yes) = [P (weight=no | Diabetes=yes) * P (Sleep=yes | Diabetes=yes) * P (water=yes | Diabetes=yes) * P (Heat=yes | Diabetes=yes) * P (Hungry=yes | Diabetes=yes)] * P (Diabetes=yes)

P (Diabetes=yes) = 0.4 * 0.95 * 0.7 * 0.85 * 0.6 * 0.66

= 0.0895

≈ 0.1

P (Diabetes=no) = [P (weight=no | Diabetes=no) * P (Sleep=yes | Diabetes=no) * P (water=yes | Diabetes=no) * P (Heat=yes | Diabetes=no) * P (Hungry=yes | Diabetes=no)] * P (Diabetes=no)

P (Diabetes=no) = 0.6 * 0.7 * 0.6 * 0.7 * 0.4 * 0.33

= 0.0232

Since the probability of Diabetes = yes is more than that for Diabetes = no, hence it can be concluded that this particular prospect is more probable to be diabetic.

Similarly, computations are performed for other newly discovered twitter results to predict if the prospect is diabetic or not.

4.2 RANDOM FOREST METHOD OF CLASSIFICATION

Random Forest method of classification was used to create a pattern recognition and prediction on the basis of the trained data. Following steps led to creation of random forest architecture using Matlab:

Firstly, the training set of data is taken into account by the algorithm.

- Clustering of the trained data sets into groups and subgroups is done and the structure would look similar to that of a tree called a decision tree. Clusters at each node are chosen randomly by the program to judge the relationship between the data points.
- A forest is formed by counting multiple trees and each tree is different in the forest because of the randomness of the variables chosen.
- The dataset except for the training dataset is used to classify the data points for the new dataset.
- The tree which the maximum prediction number is considered and shown as the output by the random forest algorithm [53].

The results obtained after running the Matlab code for Random forest is a plot of out-of-bag error over the number of grown classification trees, as shown in Figure 4.1. As the number of trees grown increases the out-of-bag error should decrease, if the result is close to accurate.

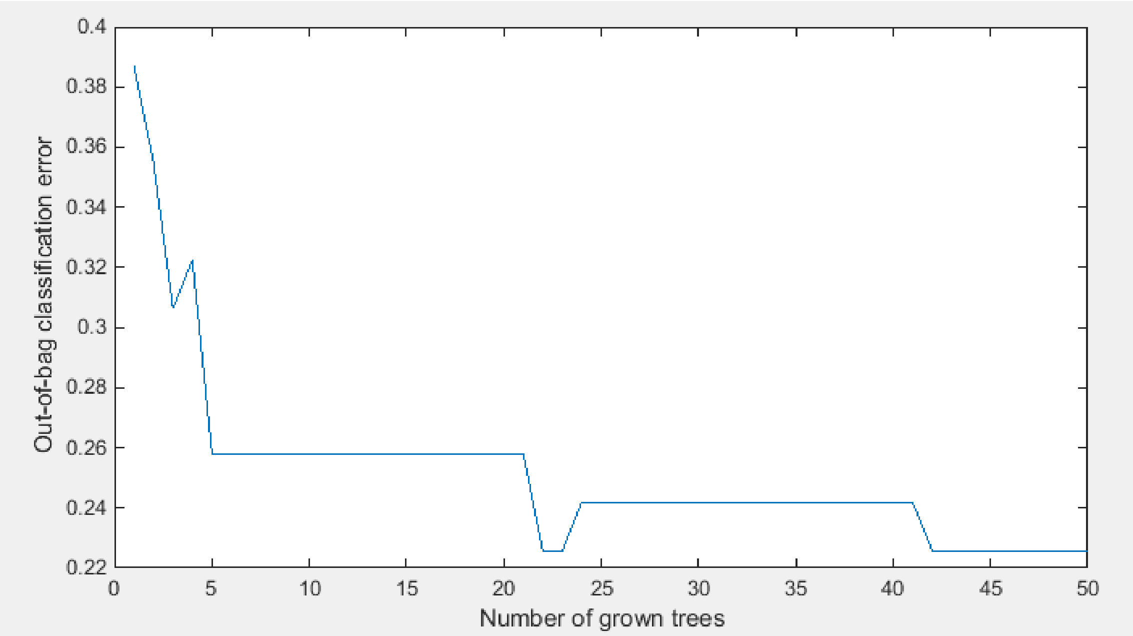


Figure. 4.1. Out-of-bag v/s Number Of Trees Grown Plot

From the plot obtained it can be seen that the out-of-bag error eventually decreases with the number of trees grown depicting the fall in generalization error.

5. CONCLUSION

The study began with the following premise:

Is it possible to observe diabetes based on text analysis of social media even if the individual does not intentionally discuss his/her health?

As evidenced by the results provided in the study, a non-pandemic disease like diabetes can be diagnosed in its early stages by analyzing the posts on social networking sites. There have been other approaches to diagnose a non-pandemic or a non-seasonal disease before, however, they have been developed in medical settings. This raised a question if such diseases can be diagnosed without any physical check-up of the person and without spending time waiting for the disease to aggravate. While the focus has been laid only on diabetes in general, and not its two categories – Type 1 and Type 2, these methods can also be applied to diseases such as cancer, which has been one of the leading causes of death among people due to the lack of prior knowledge. This could result in serious social and economic effects. Since people now-a-days have become more acquainted to social media and prefer discussing about their on-goings with everyone they know and are also open about sharing their thoughts and opinions with others looking for some healthy discussions, these methods of diagnosis have henceforth become more convenient and accurate. In conclusion, the study demonstrates the feasibility of diagnosing diabetes on the basis of text analysis.

6. FUTURE WORK

The work presented in this thesis is barely the tip of an iceberg. This section describes a few approaches that can potentially be used to improve the classification accuracy. It is intended as a guide for future researchers who wish to extend and build on this thesis. Few of the avenues for improvement include:

- **Larger Data Samples:** The first and most obvious improvement of all is to collect more number of data. In this thesis, a total of 30 Twitter IDs were considered. One of the immediate tasks is to increase the number of samples and reassess the reliability of the framework.
- **Additional Control Factors:** Data can be collected from other sites, for example Facebook, Instagram, etc., as well in order to assess the pattern. With additional attributes such as geographic locations, ethnicity, age etc., more fine grained statistical insights can be gained.
- **Automated Feature Extraction:** As feature extraction is crucial for supervised learning models, it might be beneficial to investigate the possibility of automating the feature extraction process using sentiment analysis. Use of natural language processing and computational linguistics to extract subjective information in source materials would reduce the human involvement for analyzing the tweets and hence provide homogenous results.

APPENDIX A

JAVA CODE TO COUNT THE WORDS

```
package diabetesData;

import java.io.BufferedReader;

import java.io.DataInputStream;

import java.io.FileInputStream;

import java.io.FileNotFoundException;

import java.io.IOException;

import java.io.InputStreamReader;

import java.util.ArrayList;

import java.util.Collections;

import java.util.Comparator;

import java.util.HashMap;

import java.util.List;

import java.util.Map;

import java.util.Set;

import java.util.StringTokenizer;

import java.util.Map.Entry;


public class MaxDuplicateWordCount {


public Map<String, Integer> getWordCount(String fileName){


FileInputStream fis = null;

DataInputStream dis = null;
```

```
BufferedReader br = null;
Map<String, Integer> wordMap = new HashMap<String, Integer>();
try {
fis = new FileInputStream(fileName);
dis = new DataInputStream(fis);
br = new BufferedReader(new InputStreamReader(dis));
String line = null;
while((line = br.readLine()) != null){
StringTokenizer st = new StringTokenizer(line, " ");
while(st.hasMoreTokens()){
String tmp = st.nextToken().toLowerCase();
if(wordMap.containsKey(tmp)){
wordMap.put(tmp, wordMap.get(tmp)+1);
} else {
wordMap.put(tmp, 1);
}
}
}
} catch (FileNotFoundException e) {
e.printStackTrace();
} catch (IOException e) {
e.printStackTrace();
} finally{
```

```
try{if(br != null) br.close();}catch(Exception ex){}
}
return wordMap;
}

public List<Entry<String, Integer>> sortByValue(Map<String, Integer>
wordMap){

Set<Entry<String, Integer>> set = wordMap.entrySet();
List<Entry<String, Integer>> list = new ArrayList<Entry<String,
Integer>>(set);
Collections.sort( list, new Comparator<Map.Entry<String, Integer>>()
{
public int compare( Map.Entry<String, Integer> o1, Map.Entry<String,
Integer> o2 )
{
return (o2.getValue()).compareTo( o1.getValue() );
}
} );
return list;
}
public static void main(String a[]){
MaxDuplicateWordCount mdc = new MaxDuplicateWordCount();
```

```
Map<String, Integer> wordMap =
mdc.getWordCount("C:/TheDiabeticDr2479888794.txt");
List<Entry<String, Integer>> list = mdc.sortByValue(wordMap);
for(Map.Entry<String, Integer> entry:list){
System.out.println(entry.getKey()+" ==== "+entry.getValue());
}
}
```

APPENDIX B

RAW DATA USED FOR THE FISHER'S EXACT TEST

Month	*Diabetes*	*Sleep*	*Water*	*Rash*	*Tired*
Jan-09	66	66	53	55	59
Feb-09	71	66	55	55	67
Mar-09	76	69	57	57	76
Apr-09	72	65	60	61	81
May-09	71	67	59	66	79
Jun-09	67	64	62	68	78
Jul-09	64	67	61	70	77
Aug-09	64	68	60	67	73
Sep-09	66	66	54	59	70
Oct-09	67	68	52	55	66
Nov-09	75	67	51	54	70
Dec-09	61	66	47	53	60
Jan-10	66	78	52	53	68
Feb-10	70	73	54	55	73
Mar-10	73	73	56	56	77
Apr-10	71	71	57	63	85
May-10	67	70	60	67	79
Jun-10	65	66	62	75	87
Jul-10	62	75	67	81	85
Aug-10	63	78	62	73	79
Sep-10	66	86	57	64	82
Oct-10	67	79	54	61	71
Nov-10	73	75	51	55	73
Dec-10	58	75	51	58	62
Jan-11	66	82	55	60	75
Feb-11	66	76	55	61	73
Mar-11	69	77	57	62	86
Apr-11	67	78	61	69	95
May-11	66	80	64	70	93
Jun-11	61	83	68	79	100
Jul-11	61	83	68	84	94
Aug-11	61	83	63	71	81
Sep-11	67	82	57	68	74
Oct-11	67	81	55	64	68
Nov-11	75	83	54	64	67
Dec-11	61	80	50	61	61
Jan-12	57	87	56	66	74
Feb-12	70	87	58	69	82
Mar-12	72	84	58	73	86
Apr-12	68	83	61	75	92
May-12	70	82	66	85	90
Jun-12	63	83	66	84	85

Jul-12	63	86	70	92	83
Aug-12	65	89	67	87	82
Sep-12	66	87	60	74	76
Oct-12	66	84	56	66	72
Nov-12	71	83	54	65	66
Dec-12	56	82	51	65	61
Jan-13	63	91	59	69	69
Feb-13	68	86	59	67	75
Mar-13	68	90	60	72	79
Apr-13	71	86	63	79	83
May-13	68	85	66	83	82
Jun-13	66	88	69	88	79
Jul-13	64	89	72	94	74
Aug-13	64	95	68	87	77
Sep-13	67	93	62	75	70
Oct-13	71	92	58	69	72
Nov-13	75	91	58	68	67
Dec-13	60	90	56	67	57

APPENDIX C

MATLAB CODE USED FOR RANDOM FOREST CLASSIFICATION

```
% Since TreeBagger uses randomness different results can be expected each
% time the program is run.
% This makes sure that the results obtained are same every time the code is run.
rng default
BaggedEnsemble = TreeBagger(60,cali,classLabels,'OOBPred','On')
% Here some training data is created.
% The rows&lt; represent the samples or individuals.
% The first two columns represent the individual's features.
% The last column represents the class label
trainData = [ ...
[1,  1,  0,  1,  1,  1];
[1,  1,  1,  1,  1,  1];
[1,  1,  1,  0,  1,  1];
[1,  1,  1,  0,  1,  1];
[1,  1,  1,  1,  1,  1];
[1,  1,  0,  0,  1,  1];
[1,  1,  1,  1,  1,  1];
[0,  0,  0,  0,  0,  0];
[1,  1,  0,  1,  1,  1];
[1,  1,  0,  1,  1,  1];
[1,  0,  0,  0,  1,  0];
[1,  1,  1,  1,  1,  1];
[0,  0,  0,  0,  0,  0];
[1,  1,  1,  1,  1,  1];
[1,  1,  1,  1,  1,  1];
[1,  1,  1,  0,  1,  1];
[1,  1,  0,  0,  0,  1];
[1,  1,  0,  1,  1,  0];
[1,  1,  1,  1,  1,  0];
[1,  1,  1,  1,  1,  0];
[1,  1,  1,  1,  1,  1];
[1,  1,  1,  1,  1,  0];
[0,  1,  0,  0,  0,  1];
[1,  0,  0,  0,  0,  1];
[1,  0,  0,  0,  1,  0];
[1,  1,  0,  0,  1,  0];
[1,  0,  0,  0,  1,  0];
[0,  1,  1,  0,  1,  0];
[1,  1,  1,  0,  1,  0];
[1,  0,  0,  1,  0,  0];
];
cali = trainData(:,(1:5))
classLabels = trainData(:,6)
% How many trees do you want in the forest?
nTrees = 50
% Train the TreeBagger (Decision Forest).
```

```
B = TreeBagger(nTrees,cali,classLabels, 'Method', 'classification');
% Given a new individual WITH the features and WITHOUT the class label,
newData1 = [1,  1,  1,  1,  0];
% Use the trained Decision Forest.
predChar1 = B.predict(newData1);% Predictions is a char though.
predictedClass = str2double(predChar1)
oobErrorBaggedEnsemble = oobError(BaggedEnsemble);
plot(oobErrorBaggedEnsemble)
xlabel 'Number of grown trees';
ylabel 'Out-of-bag classification error';
```

APPENDIX D

JAVA CODE TO GET USER STATUS

(By, Raja Ashok Bolla)

```
/**
 * This Class is used to get the list of status.
 */
package FarheenTweetsPack;
import java.io.BufferedReader;
import java.io.File;
import java.io.FileNotFoundException;
import java.io.FileReader;
import java.io.FileWriter;
import java.io.IOException;
import java.io.PrintWriter;
import java.util.ArrayList;
import java.util.List;
import com.data.region.trending.AllKeys;
import twitter4j.PagableResponseList;
import twitter4j.Paging;
import twitter4j.ResponseList;
import twitter4j.Status;
import twitter4j.Twitter;
import twitter4j.TwitterException;
import twitter4j.TwitterFactory;
import twitter4j.User;
import twitter4j.conf.ConfigurationBuilder;
```

```
public class GetUserStatus {
static String ckey = "BQMS6OKOPQgjhQUUo8TmXcigU";
static String cSecret =
"Kpz8CVEmllp2aQ5mXZa6vZB7jxOupVP7GrcsNs5w1q41EPQZ01";
static String tKey = "282016016-
o1dfgiPLYWUepFAqnUK1ZZY4EHlQuZ3KzxD9IdAL";
static String tSecret = "FYUlc73WlO7TeGFDpo6oj68KdVQRdLrzYRhps7kWsPcFk";
static Twitter twitter;
@SuppressWarnings("unchecked")
public static void main(String[] args) throws IOException, TwitterException {
ConfigurationBuilder cb = new ConfigurationBuilder();
cb.setDebugEnabled(true).setOAuthConsumerKey(ckey)
.setOAuthConsumerSecret(cSecret).setOAuthAccessToken(tKey)
.setOAuthAccessTokenSecret(tSecret);
// FileWriter outFile1 = new FileWriter("Tweets123.txt", true);
// try {
TwitterFactory factory = new TwitterFactory(cb.build());
twitter = factory.getInstance();
ResponseList<User> users = null;
String[] srch_ids = loadUserIDs();
int count = 0;
for (String s : srch_ids) {
if (s == null)
```

```
System.exit(0);

String[] srch = new String[] { s };

try {

users = twitter.lookupUsers(srch);

} catch (TwitterException tee) {

if (tee.toString().contains("Could not authenticate you")) {

System.out

.println("##################Junk ID########################"

+ s);

try {

// Introduced delay of 15 minutes due to Twitter

// Limitations

Thread.sleep(15 * 60 * 1000);

} catch (InterruptedException e1) {

// TODO Auto-generated catch block

e1.printStackTrace();

}

users = twitter.lookupUsers(srch);

}

}

for (User user : users) {

String uName = user.getScreenName().toString();

// Folder to store the extracted tweets
// Username is generally the name of the file

File f = new File("FarheenTweets/" + uName + ".txt");

FileWriter outFile1 = new FileWriter(f, true);

PrintWriter out1 = new PrintWriter(outFile1);

// System.out.println(user.getName() + "  :  ");

long cursor = -1;

Paging paging = new Paging(1);

ArrayList<Status> tweets = null;

// PagableResponseList<User> followers;

search: do {

System.out.println(count + " : " + uName);

count++;

try {

tweets = (ArrayList<Status>) twitter.getUserTimeline(

uName, paging);

} catch (Exception e) {

if (e.toString().contains("Rate limit exceeded")) {

// Handling of Rate Limit Exception

try {

Thread.sleep(15 * 60 * 1000);

} catch (InterruptedException e1) {

// TODO Auto-generated catch block

e1.printStackTrace();
}
try {
tweets = (ArrayList<Status>) twitter
.getUserTimeline(uName, paging);
} catch (Exception e1) {
if (e1.toString().contains(
"Rate limit exceeded")) {
try {
Thread.sleep(15 * 60 * 1000);
} catch (InterruptedException e2) {
// TODO Auto-generated catch block
e2.printStackTrace();
}
} else
{
break search;
}
}
} else {
break search;
}
}
if (tweets == null)
```

```
break search;
for (Status message : tweets) {
// Writing the timestamped tweets in the file
out1.write(message.getCreatedAt() + " Msg : "
+ message.getText());
out1.write("\n");
}
paging.setPage(paging.getPage() + 1);
} while (tweets.size() > 0 && paging.getPage() < 40);
out1.close();
}
}
}
private static String[] loadUserIDs() {
String[] ids = new String[50912];
int count = 0;
FileReader fr;
try {
// A place where application looks up for names of the users to
// search the tweets
fr = new FileReader(new File("Training//farheenids.txt"));
BufferedReader br = new BufferedReader(fr);
String thisLine;
```

```
String[] toks;

while ((thisLine = br.readLine()) != null) {

ids[count] = thisLine;

count++;

}

} catch (FileNotFoundException e) {

e.printStackTrace();

} catch (IOException e) {

e.printStackTrace();

}

return ids;

}

}
```

BIBLIOGRAPHY

[1] Sun’s J. Kelly. Diabetes. Centers for disease control and prevention, page – 1.

[2] Diabetes basic symptoms. In Diabetes, Org. Retrieved December 15, 2014, from http://www.diabetes.org/diabetes-basics/symptoms/.htm.

[3] T. Bodnar and M. Salath ́e. Validating models for disease detection using twitter. In Proceedings of the 22nd International Conference on World Wide Web Companion, WWW ’13 Companion, pages 699–702, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.

[4] D. Butler. When Google got flu wrong. Nature, 494(7436):155–156, Feb. 2013.

[5] B. Milne. Crossing the line on social media? (And we don’t mean compliance.). Retrieven March 7, 2015, from http://socialware.com/2015/02/04/crossing-line-social-media-dont-mean-complianc/.htm.

[6] Twitter. In Wikipedia. Retrieved March 7, 2015, from http://en.wikipedia.org/wiki/Twitter.htm.

[7] Boyd DM, Ellison NB. Social network sites: Definition, history, and scholarship. J Comp Med Commun.2008; 13:210–230, from http://onlinelibrary.wiley.com/doi/10.1111/j.1083-6101.2007.00393.x/abstract;jsessionid=A99355C656252603F1643799F7F0E730.f02t02.htm.

[8] Thackeray R, Neiger BL, Hanson CL, McKenzie JF. Enhancing promotional strategies within social marketing programs: use of Web 2.0 social media. Health Promot Pract. 2008 Oct; 9(4):338–43, from http://hpp.sagepub.com/content/9/4/338.htm.

[9] Kaplan AM, Haenlein M. Users of the world, unite! The challenges and opportunities of social media.Business Horizons. 2010; 53:59–68.

[10] Maness JM. Library 2.0 Theory: Web 2.0 and its implications for libraries. 2006, from http://www.webology.org/2006/v3n2/a25.html.

[11] Kamel Boulos MN, Wheeler S. The emerging Web 2.0 social software: an enabling suite of sociable technologies in health and health care education. Health Info Libr J. 2007 Mar; 24(1):2–23, from http://onlinelibrary.wiley.com/doi/10.1111/j.14711842.2007.00701.x/abstract.

[12] Correa T, Willard Hinsley A, de Zúñiga HG. Who interacts on the Web? The intersection of users' personality and social media use. Computers in Human Behavior. 2010; 26(2):247–253.

[13] Dawson J. Doctors join patients in going online for health information. New Media Age. 2010; 7.

[14] Young SD, Rice E. Online social networking technologies, HIV knowledge, and sexual risk and testing behaviors among homeless youth. AIDS Behav. 2011 Feb; 15(2):253–60, from http://europepmc.org/abstract/MED/20848305.htm.

[15 Hanson C, West J, Neiger B, Thackeray R, Barnes M, McIntyre E. Use and acceptance of social media among health educators. Am J Health Educ. 2011; 42(4):197–204.

[16] Dowdell EB, Burgess AW, Flores JR. Original research: online social networking patterns among adolescents, young adults, and sexual offenders. Am J Nurs. 2011 Jul; 111(7):28–36; quiz 37, from http://journals.lww.com/ajnonline/pages/articleviewer.aspx?year=2011&issue=07000&article=00021&type=abstract.htm.

[17] Cobb NK, Graham AL, Abrams DB. Social network structure of a large online community for smoking cessation. Am J Public Health. 2010; 100(7):1282–1289, from http://ajph.aphapublications.org/doi/abs/10.2105/AJPH.2009.165449.htm.

[18] Kontos EZ, Emmons KM, Puleo E, Viswanath K. Communication inequalities and public health implications of adult social networking site use in the United States. J Health Commun. 2010 Dec; 15 Suppl 3:216–35, from http://europepmc.org/abstract/MED/21154095.htm.

[19] Lariscy RW, Reber BH, Paek H. Examination of Media Channels and Types as Health Information Sources for Adolescents: Comparisons for Black/White, Male/Female, Urban/Rural. Journal of Broadcasting & Electronic Media. 2010 Mar 2010; 54(1):102–120, from http://www.tandfonline.com/doi/abs/10.1080/08838150903550444#.VQ3VRPnF_y0.htm.

[20] Chew C, Eysenbach G. Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. PLoS One. 2010 Nov; 5(11):e14118, from http://dx.plos.org/10.1371/journal.pone.0014118.htm.

[21] Colineau N, Paris C. Talking about your health to strangers: understanding the use of online social networks by patients. New Review of Hypermedia and Multimedia. 2010 Apr 2010; 16(1-2):141–160, from http://www.tandfonline.com/doi/abs/10.1080/13614568.2010.496131#.VQ3VyvnF_y0.htm.

[22] Heidelberger CA. Health Care Professionals' Use of Online Social Networks. 2011, from *webcite*http://cahdsu.wordpress.com/2011/04/07/infs-892-health-care-professionals-use-of-online-social-networks/.

[23] Chou WY, Hunt Y, Folkers A, Augustson E. Cancer survivorship in the age of YouTube and social media: a narrative analysis. J Med Internet Res. 2011 Jan; 13(1):e7, from http://www.jmir.org/2011/1/e7/.htm.

[24] Heidelberger CA. Health Care Professionals' Use of Online Social Networks. 2011, from http://cahdsu.wordpress.com/2011/04/07/infs-892-health-care-professionals-use-of-online-social-networks/.htm.

[25] Fox S, Jones S. The Social Life of Health Information. 2009, from http://www.pewinternet.org.Reports/2007/Information-Searches.htm.

[26] McNab C. What social media offers to health professionals and citizens? 2009, from http://www.who.int/bulletin/volumes/87/8/09-066712/en/.htm.

[27] Eyrich N, Padman ML, Sweetser DS. PR practitioners' use of social media tools and communication technology. Public Relations Review. 2008; 34:412–414.

[28] Green B, Hope A. Promoting clinical competence using social media. Nurse Educ. 2010; 35(3):127–9, from http://journals.lww.com/nurseeducatoronline/pages/articleviewer.aspx?year=2010&issue=05000&article=00015&type=abstract.htm.

[29] Giustini D. How Web 2.0 is changing medicine. BMJ. 2006 Dec 23; 333(7582):1283–4, from http://europepmc.org/abstract/MED/17185707.htm.

[30] Bosslet GT, Torke AM, Hickman SE, Terry CL, Helft PR. The patient-doctor relationship and online social networks: results of a national survey. J Gen Intern Med. 2011 Oct; 26(10):1168–74, from http://europepmc.org/abstract/MED/21706268.htm.

[31] Frost JH, Massagli MP. Social uses of personal health information within PatientsLikeMe, an online patient community: what can happen when patients have access to one another's data? J Med Internet Res.2008 May; 10(3):e15, from http://www.jmir.org/2008/3/e15/.htm.

[32] Wicks P, Massagli M, Frost J, Brownstein C, Okun S, Vaughan T, Bradley R, Heywood J. Sharing health data for better outcomes on PatientsLikeMe. J Med Internet Res. 2010 Jun; 12(2):e19, from http://www.jmir.org/2010/2/e19/.htm.

[33] Farmer AD, Bruckner Holt CE, Cook MJ, Hearing SD. Social networking sites: a novel portal for communication. Postgrad Med J. 2009 Sep; 85(1007):455–9, from http://pmj.bmj.com/content/85/1007/455.htm.

[34] Adams SA. Blog-based applications and health information: two case studies that illustrate important questions for Consumer Health Informatics (CHI) research. Int J Med Inform. 2010 Jun; 79(6):e89–96, from http://www.ijmijournal.com/article/S1386-5056(08)00103-2/abstract.htm.

[35] Lagu T, Kaufman EJ, Asch DA, Armstrong K. Content of weblogs written by health professionals. J Gen Intern Med. 2008 Oct; 23(10):1642–6, from http://europepmc.org/abstract/MED/18649110.htm.

[36] Versteeg KM, Knopf JM, Posluszny S, Vockell AL, Britto MT. Teenagers wanting medical advice: Is MySpace the answer? Arch Pediatr Adolesc Med. 2009 Jan; 163(1):91–2, from http://europepmc.org/abstract/MED/19124711.htm.

[37] Greene JA, Choudhry NK, Kilabuk E, Shrank WH. Online social networking by patients with diabetes: a qualitative evaluation of communication with Facebook. J Gen Intern Med. 2011 Mar; 26(3):287–92, from http://europepmc.org/abstract/MED/20945113.htm.

[38] Lagu T, Hannon NS, Rothberg MB, Lindenauer PK. Patients' evaluations of health care providers in the era of social networking: an analysis of physician-rating websites. J Gen Intern Med. 2010 Sep; 25(9):942–6, http://europepmc.org/abstract/MED/20464523.htm.

[39] Signorini A, Segre AM, Polgreen PM. The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. PLoS One. 2011 May; 6(5):e19467, from http://dx.plos.org/10.1371/journal.pone.0019467.htm.

[40] Salathé M, Khandelwal S. Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control. PLoS Comput Biol. 2011 Oct; 7(10):e1002199, from http://dx.plos.org/10.1371/journal.pcbi.1002199.htm.

[41] Corley CD, Cook DJ, Mikler AR, Singh KP. Text and structural data mining of influenza mentions in Web and social media. Int J Environ Res Public Health. 2010 Feb; 7(2):596–615, from http://www.mdpi.com/1660-4601/7/2/596.htm.

[42] J. V. Freeman, M. J. Campbell, THE ANALYSIS OF CATEGORICAL DATA: FISHER’S EXACT TEST. Retrieved March 5, 2015, from http://www.sheffield.ac.uk/polopoly_fs/1.43998!/file/tutorial-9-fishers.pdf.

[43] D. R. Olson, K. J. Konty, M. Paladini, C. Viboud, and L. Simonsen. Reassessing Google Flu Trends Data for Detection of Seasonal and Pandemic Influenza: A Comparative Epidemiological Study at Three Geographic Scales. PLoS computational biology, Oct. 2013. p. 1-2.

[44] T. Bodnar, V. C. Barclay, N. Ram, C. S. Tucker, M. Salathe. On the Ground Validation of Online Diagnosis with Twitter and Medical Records. International World Wide Web Conference Committee (IW3C2), Apr. 2014. Retrieved December 15, 2014, from http://arxiv-web3.library.cornell.edu/pdf/1404.3026v1.pdf.htm.

[45] Diabetes and summer. In Mayoclinic, Org. Retrieved December 15, 2014, from http://www.mayoclinic.org/diseases-conditions/diabetes/expert-blog/diabetes-and-summer/bgp-20056545.htm.

[46] Fisher's exact test. In Wikipedia. Retrieved December 15, 2014, from http://en.wikipedia.org/wiki/Fisher%27s_exact_test.htm.

[47] Naive Bayes Classifier. In Wikipedia. Retrieved December 15, 2014, from http://en.wikipedia.org/wiki/Naive_Bayes_classifier.htm.

[48] Naive Bayes classifier. In Creative Commons, Org. Retrieved March 2, 2015, from http://www.ic.unicamp.br/~rocha/teaching/2011s2/mc906/aulas/naive-bayes-classifier.pdf, p-1.

[49] Naive Bayes Classifier. In Wikipedia. Retrieved December 15, 2014, from http://www.saedsayad.com/naive_bayesian.htm.

[50] Random forest. In Wikipedia. Retrieved December 15, 2014, from http://en.wikipedia.org/wiki/Random_forest.htm.

[51] A. Liaw, M. Wiener, Classification and Regression by randomForest, in vol. 2/3, Dec, 2012 from http://ftp3.ie.freebsd.org/pub/download.sourceforge.net/pub/sourceforge/i/ii/iiitbprj1/LiteratureSurvey/Liaw_02_Classification%20and%20regression%20by%20randomForest.pdf. P-1.

[52] A. L. Boulesteix, S. Janitza, J. Kruppa, I. R. Konig. Overview of Random Forest Methodology and Practical Guidance with Emphasis on Computational Biology and Bioinformatics, July 25, 2012 from http://epub.ub.uni-muenchen.de/13766/1/TR.pdf. P – 2-3.

[53] Random forest. In Stat Berkeley, Edu. Retrieved December 15, 2014, from https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm.

VITA

Farheen Ali was born in Orissa, India. In August 2012, she received her Bachelor’s degree in Computer Science from Centurion Institute of Technology, India. She then worked as a Business Analyst for a year with Gram Tarang Employability Training Services, India, untill July 2013. She subsequently joined Missouri University of Science and Technology (formerly University of Missouri – Rolla) in Fall 2013. She completed her Master’s degree in Information Science and Technology and earned a Graduate Certificate in Business Intelligence in May 2015. During the course of her Master’s degree she pursued co-op term with Sysintelli Inc. in 2015.