SUDAN UNIVERSITY OF SCIENCE AND TECHNOLOGY

COLLEGE OF GRADUATE STUDIES

Building a Classification Model for Diseases Discovery from Tweets

بناء نموذج تصنيفي لاكتشاف الأمراض من التغريدات

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS OF MSC. IN INFORMATION TECHNOLOGY

BY

Mojahed Salih Mohammed Ibrahim

Supervised by

Dr. Nisreen Beshir Osman

December 2018

الآية

(ألم تَرَ أنَّ الله أنزلَ من السَّماء ماءً فأخرجنا به ثمراتٍ مُخْتلفاً ألوانُها ومن الجبالِ جُدَدٌ بيضٌ وحمرٌ مختلفٌ ألوانُها وغرابيبُ سُودٌ ومن النَّاسِ والدَّوابِّ والأنعامِ مُختلفٌ ألوانُهُ كذلك إنَّما يَخشى الله من عبادهِ العلماءُ إنَّ الله عزيزٌ غفور)

(35 فاطر آية 27-28).

Dedication

I dedicate this thesis to my parents, Saleh Mohammed Ibrahim and Nemat Mohamed Adam who supported me to continue my educational process, to my colleagues and finally to my teacherswho gave me advices that helped me in my thesis. I will be grateful forever for your love.

Acknowledgment

I would like to express my gratitude to my supervisor Dr. Nisreen Beshir for the useful comments, remarks and engagement through the learning process of this master thesis. Furthermore I would like to thank Merghani Mohammed for introducing me to the topic as well for the support on the way. I would like to thank my loved ones, who have supported me throughout entire process, both by keeping me harmonious and helping me putting pieces together.

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 building a classification model that could use social media (Twitter) as the data source for prediction diseases name based on text mining of tweets.

The aim of this research is to building a classification model at the individual level using a sample of tweets that have been collected from the Twitter API using the query including keywords: "blood pressure, bone, cancer, diabetes, kidney, lung cancer and migraines" based on text mining of tweets.

This thesis investigates the feasibility of monitor and classifying diseases via machine learning techniques, Support Vector Machine and Random Forest classifiers.The results of the classification models showed high accuracy rate in the prediction of diseases name.

المستخلص

الابتكار في التكنولوجيا يمكّن الناس من التواصل ومشاركة المعلومات والبحث عن احتياجاتهم بمجرد الجلوس في الغرف والقيام ببعض النقرات. بينما لعبت وسائل التواصل الاجتماعي دورًا مهمًا للغاية في تواصل الأشخاص حول العالم ، إلا أن إمكاناتها تجاوزت الفكرة الابتكارية المتمثلة في ربط الأشخاص عبر شبكاتهم الاجتماعية. بينما يعتقد الكثيرون أنه لا توجد نقطة إلتقاء لقطاع الرعاية الصحية ووسائل التواصل الاجتماعي ، فقد كانت مفاجأة عندما أظهرت الأبحاث والابتكارات أن وسائل التواصل الاجتماعي يمكن أن تلعب دورًا مهمًا للغاية في قطاع الرعاية الصحية.

تم إجراء البحث في بناء نموذج تصنيف يمكن أن يستخدم وسائل التواصل الاجتماعي مثال (Twitter) كمصدر للبيانات الخاصة بالامراض التنبؤ بناءً على التنقيب في نص التغريدات.

الهدف من هذا البحث هو بناء نموذج تصنيف على المستوى الفردي باستخدام عينة من التغريدات التي تم جمعها من واجهة برمجة تطبيقات Twitter باستخدام الاستعلام بما في ذلك الكلمات الرئيسية: "ضغط الدم والعظام والسرطان والسكري والكلى وسرطان الرئة والصداع النصفي "على أساس التعدين النص من تويت.

تبحث هذه الأطروحة في جدوى رصد وتصنيف الأمراض من خلال تقنيات التعلم الآلي ومصنفات ناقلات الدعم ومصنفات الغابات العشوائية. أظهرت نتائج نماذج التصنيف نسبة دقة عالية في التنبؤ باسم الأمراض.

Table of Contents

Table of Abbreviations

Abbreviation	Terms
API	Application Programming Interface
REST	Representational State Transfer
CRUD	Create Retrieve Update Delete
FDA	Food and Drug Administration
EI	Epidemic Intelligence
SVM	Support Vector Machine
ILI	Influenza and Influenza
CDC	Control and Prevention
ML	Machine Learning

Chapter One: Introduction

1.1 Overview

Nowadays, social media is been used by many researchers to collect data on Patient experiences and opinions such as symptoms, physician's performance etc. with the help of these new modes of interactions, social media can monitor public response to health issues, track and monitor disease outbreaks, identify target areas for intervention efforts, and disseminate pertinent health information to targeted communities [1]. Health professionals can aggregate data about patient experiences from blogs and monitor the public reaction to health issues.

1.2 Problem Statement

Predicting the disease through social media is very important for decision makers in health care. Helps them make timely decisions. This is the importance of analyzing medical data from social media. One type of data that gives us basic knowledge about the disease is the medical data on these social media.

Since there are not many studies that analyze this topic there is a need to monitor diseases based on text analysis and social media.

1.3 Research Aims

The main aim of this research is prediction diseases from social media especially Twitter by building a classifier model for tweets.

In order to build the model, the following objectives needs to be achieved:

1. Data collection is a critical phase because no quality data no quality mining results.
2. Select the best algorithms that used in text mining.
3. Machine Learning helping us to achieve our aims.

1.4 Research Scope

This is research focus on building a classifier model to prediction disease name into the tweets based on text analysis of tweets.
Dataset contents 5194 tweets for 10 disease name collected through Twitter API using hashtag (#).

1.5 Research Methodology

The methodology was divided into six main steps: Twitter data collection, Data preparation, Create a document term matrix, Training data, Classifier model and Results; to easier handle methodology.

1.6 Research Organization

The structure of this research divided into five chapters as shown below:

Chapter One: Introduction This chapter describes overview and the whole idea behind the research and the problem statement, aims, scope, methodology, and organization of research. **Chapter Two: Literature Review & Related Work** Divided into two section. First one takes the major concept about topic and the second one discusses the previous studies. **Chapter Three: Research Methodology** This chapter represents the methodology of research. The contents are: the way of building classifier model from scratch. **Chapter Four: Results** discuss the results of applying the classifier model and compared the results between two algorithms. **Chapter five: Conclusions and Recommendations** finally **References and Appendices.**

Chapter Two: Literature Review & Related Work

2. 1 Introduction

This chapter is divided into two sections. The first section gives general description about Twitter, Healthcare Information on Twitter, Data Mining, Knowledge Discovery and Sentiment Analysis. The second section describes the related studies to research project.

2.2 Literature Review

2.2.1 Twitter

Twitter is an online social networking and microblogging service that enables users to send and read short 140-character text message, called "tweets". Registered users can read and post tweets, but unregistered users can only read them [2]. There are various ways for users to access Twitter, like official website interface, SMS, or mobile device app.

On Twitter, users can add a friend by searching for his or her user name. After the friend relationship is generated, users are able to start to view the updates of their friends. Another relationship is called "follower". A user becomes someone's follower when he starts following someone on Twitter, which means the user is subscribing to updates of users that he follows. This helps users to build relationship with those who share the same interests. Friends and followers together facilitate the users interactions [3] through both one-way and mutual way relationships.

Many features are provided for users to have a better communication. The "reply" button is similar to reply function in email. By clicking "reply", a user can respond to the tweets. Direct tweets can be sent to dedicated users by using "mention". Simply compose a tweet containing "@" followed by the username and the tweet will be sent as a message to the mentioned user. In addition, the mentions show up as links in the tweet. Retweet is another important content oriented interaction feature of Twitter. If a user likes the tweet and would like to share it with others, he can click the "retweet" button. After that all his followers will see the tweet in the updates. The retweeted tweet starts with RT to indicate that this is a retweeted tweet. Hashtags [4] are used to categorize tweets by keyword. People use the hashtag symbol # [5] before a relevant keyword or phrase in their tweet to categorize those Tweets and help them show more easily in Twitter Search. Clicking on a hashtagged word in any message shows you all other tweets marked with that keyword. Hashtags can occur anywhere in the tweet, at the beginning, middle, or end. Hashtagged words that become very popular are often Trending Topics.

According to Twitter statistics in 2013, there are 271 million monthly active users and 100 million daily active users. 500 tweets are posted per day. 29% users check Twitter multiple times a day. 52 million users live in US. Projected number of

Twitter users by 2018 will be 400 million [6]. Currently Twitter supports more than 35 languages.

For reliability, Twitter sets some technical limits to reduce downtime and error functions. First, the maximum length of the tweet content is limited to 140 characters including the links. Second, all text should be converted to UTF-8 before sending to twitter to avoid errors. Third, 250 direct messages are allowed per day. Forth, the daily update limit is 2400 per day including both tweets and retweets. Fifth, 1000 following times per day are allowed to prohibit aggressive following behaviour. Sixth, once an account is following 2000 other users, additional follow attempts are limited by account-specific ratios. Seventh, in version 1.1 of the API, an OAuth-enabled application could initiate 350 GET- based requests per hour per access token [7].

2.2.1.1 Twitter REST API

Representational state transfer (REST) is an abstraction of the architecture of the World Wide Web (WWW). More precisely, REST is an architectural style consisting of a coordinated set of architectural constraints applied to components, connectors, and data elements, within a distributed hypermedia system. REST ignores the details of component implementation and protocol syntax in order to focus on the roles of components, the constraints upon their interaction with other components, and their interpretation of significant data element [8].

An application programing interface (API) is a set of programing instructions and standards specifies how to access a web based software application or web tool. The REST APIs enable any interactions with HTTP, such as reading data, posting (create and update) data and deleting data. Therefore, REST implements all four CRUD (Create Retrieve Update Delete) operations by sending HTTP POST, GET, PUT and DELETE requests.

A resource is exposed via a fixed Universal Resource Identifier (URI). The consuming client of a RESTful application needs to know the persistent URI to access it. All future actions should be discoverable dynamically from hypermedia links included in the representations of the resources that are returned from that URI. A media type description is needed to define hypermedia access and specify what methods are available for the resources of that type.

Twitter is not only a useful online social tool it also provides a comprehensive array of REST APIs. Developers can use these APIs to make applications, websites, widgets, and other projects that interact with Twitter. Current version 1.1 offers three main APIs, the normal REST API, the search API and the stream API. Each of the APIs represents one facet of Twitter.

The REST APIs constitute the core of the Twitter API. It enables developers to access and manipulate all of Twitter main data including timelines, status updates, and user profiles. Timelines on Twitter are collections of Tweets, ordered with the most recent first. In addition, users can use the APIs to generate and post tweets back to Twitter, favourite certain tweets, retrieve statuses, send direct messages, retweet certain tweets.

The search API exposes a way for users and developers to look up keywords within twitter content to filter query. It will return a collection of related tweet objects matching a given query with HTTP GET method. Additionally, hashtag query is supported. This function enables users to view tweets beyond their friends or followers. Furthermore, trending topics can be discovered with the help of search API.

Stream API offers a low-latency, high-volume and near-real time access to various subsets of public and protected Twitter data. This API is only accessible to authorized users. Three main streaming products are supported: streaming API, user streams and site streams. First, streaming API returns public tweet objects matching one or more query schemas. It also supports returning a small random subset tweet objects of all public updates. Second, user streams return a stream of data dedicated to the authenticated user, and are mainly used to update to the client. Third, site streams allow multiplexing of multiple user streams over a Site Stream connection. Only a preliminary number of calls are allowed to Twitter API.

2.2.1.2 Tweet Influence

As online social networking becomes more and more popular, many studies have been done to discover valuable information from it, among which social influence has drawn a lot of attention. Influence has long been studied in many fields and the findings about influence contribute a lot in advertising and marketing. On social networks, such as Twitter, the influence refers to the ability of a user to have an effect on others or the capacity to drive action. The influence can also be interpreted as the respond of one user to the activity of another user on a social network. Similarly, studying the influence on Twitter also provides new insights in social networking.

On Twitter, a small group of users who excel in spreading information is called influencers. The common characters of influential users include a larger number of audiences, more frequent updates and higher activities. In addition, the influential tweets are more likely to be retweeted than those of others. The users who have high influence tend to gain more attention than those with low influence.

Many of them are celebrities or leaders, e.g. President Obama is ranked as No.3 on Twitter [9]. He has 44,275,975 followers and created 12,164 updates. Furthermore, he is given a score of 99 out of 100 by Klout [10], which is a famous online Twitter user ranking service. Celebrities like President Obama with a tremendous number of followers can be more effective at spreading information than others. Another example is the famous photo taken during the Oscars. Ellen DeGeneres asked other actors and actresses to take a photo and upload it to Twitter. Now the photo has been retweeted over 3 million times and becomes the most retweeted photo ever. From these examples, the most influential users show their abilities to boost the rapid diffusion of opinions, promote news quickly and disseminate the popularity of political parties. In addition, studying the influence pattern can help people have a better understanding of trending flows.

How to come up with a proper approach to characterize or quantify the influence [11] on Twitter becomes an issue. Many theories have been applied to study the influence. Traditional view focuses on the influential users and regardless the role of ordinary users. In contrast, the modern theory states the users are more likely to be affected by their peers. There are both advantages and disadvantages on each theory.

Direct links [12], e.g. follower and friend relationship, represent the way information flows on social networks. Thus, in general the number of followers and friends is an important indicator of user influence. A review of MIT
Technology [13] compares three different ways to spot the most influential spreaders based on the number of followers, degree, PageRank algorithm and K-core. After comparing the advantages and disadvantages of each method, the author draws the conclusion that the sum of the number followers of each direct follower of a user would be the best way to predict the most influential spreader. However this work has its own limitation. The number of followers for each direct follower must be known which does not suit every case. In addition, to predicate how widely the information would spread based on the larger number of followers and friends is biased. To get the measure [14] of the influence, many other factors should also be taken into account.

The retweet times indicate the quality of the content and the pass-along value. The more times a tweet got retweeted, the more value it will have. In addition to retweet influence, the frequency of updates is also a significant factor. The frequency of updates points whether a user is active or passive. Followers tend to lose interests in those less active. Moreover, the use of hashtag could also add value to the influence. In general, the hashtag is used to specify certain topic or keyword in a tweet. It will gain more attention compared to other words. Now hashtag becomes more and more popular due to the ease of use. Besides, the number of mentions represents the value of the user name.

Deciding the factors is just the first step towards creating an approach to measure the tweet influence. After that, the proper weight for each factor should be determined and coordinated in order to achieve a comprehensive ranking. Each defined weight indicates a different importance of the factor while composing the tweet influence.

2.2.2 Healthcare Information on Twitter

Within 140 characters, users can tweet whatever they like. Topics range from political opinions, comments on news events, daily life to healthcare.
According to USNEWS [15], more and more medical professionals like doctors and nurses adopt social network tools like Twitter to monitor and interact with the patients. Physicians could be friends with patients online which is good for maintaining a robust relationship between them. Communication [16] via Twitter provides an alternative engagement beyond doctors' offices or hospitals. Social network has played a significant role in changing the nature and the way of health care interaction between health care organizations and consumers.

Recently, many health care organizations have established official social network accounts, for example, US Food and Drug Administration (FDA) has multiple Twitter accounts to disseminate information. Whole Foods Market also uses Twitter to reach the consumers to promote new products and answer questions. A research shows, 90% of users from 18 to 24 years indicated that the medical information shared on social network is trustworthy [15]. Besides, lots of users are reported to use Internet including social network to seek health care information. Therefore, it is important for health care related organizations and individuals to maintain public reputation [17].

Everyday large amount of health related data are transmitted on Twitter. Users post about their own health experience, reviews of treatment, medications, hospitals or doctors, and symptoms. They seek for help as well as related tips, photos and videos. Sometimes patients may find out that they receive the same advice from doctors as from social network.

Effective cost and wide reach have made Twitter a new platform for health care information exchange. With huge amount of health care related data generate every day, Twitter evolves into a potential source pool for health care. It can be used as a complementary source in addition to formal health care data. One obvious advantage is that the tweets are real time and more relevant to current trending topics. Some studies have been conducted to utilize Twitter data along with Epidemic Intelligence (EI) to analyse potential diseases outbreak [18]. A pilot study is gathering tweets including keywords relating to "flu" to analyse the trending disease activity.

Individual health related tweet might only provide limited informative value. However the aggregation of millions of tweet is large enough to provide some insights [19]. Moreover, the tweets are not separated events. Due to the created time and geography, many tweets are related to the same topic.

2.2.3 Data Mining and Knowledge Discovery

In general, data mining or knowledge discovery is a powerful new technology which refers to the process of extracting or discovering hidden insights [20] and meaning from numerous sets of data beyond simple analysis [21]. In contrast to traditional statistical methods, complicated mathematical algorithms are applied to rapidly discover the patterns in data corpus and predict the probability of occurrences in the future. It provides powerful abilities for users to predict trends, analyse behaviours, make knowledge driven decisions and cluster large amount of data. Nowadays, data mining is applicable to many fields, such as marketing, finance, communication as well as social networks.

One important prerequisite for data mining is massive data collection. With advancements in the capacity of storage [22], it becomes easier to store data in either distributed or centralized data storage. Data warehouse is a new technology, which aims to store, maintain and retrieve data. It has played a significant role in data mining for its ability of maximizing the efficiency in data accessing and analysis. A

wide range of companies have deployed and maintained large data warehouses for data mining.

Data mining involves many knowledge and techniques. Among those modelling is the key to the process of data mining. Modelling refers to the act of building a model by applying certain algorithms to a specific dataset [23]. After that the model can be applied to new dataset in another situation for automatic discovery or trend prediction. Data mining is becoming increasingly popular because it helps to providing valuable insights to the data and it can be applied to various fields.

2.2.4 Sentiment Analysis

Sentiment Analysis is a technique widely used in text mining. Twitter Sentiment Analysis, therefore means, using advanced text mining techniques to analyse the sentiment of the text (here, tweet) in the form of positive, negative and neutral [24]. Twitter Sentiment Analysis, also known as Opinion Mining, is primarily for analysing conversations, opinions, and sharing of views (all in the form of tweets) for deciding business strategy, political analysis, and also for assessing public actions.

Enginuity, Revealed Context, Steamcrab, MeaningCloud, and SocialMention are some of the well-known tools used for Twitter Sentiment Analysis. R and Python are widely used for sentiment analysis dataset twitter.

Our discussion will include, Twitter Sentiment Analysis in R and also throw light on Twitter Sentiment Analysis techniques and teach you how to generate Twitter Sentiment Analysis project report.

2.2.4.1What is the Twitter Sentiment Analysis?

Sentiment Analysis is a technique used in text mining. Twitter Sentiment Analysis may, therefore, be described as a text mining technique for analyzing the underlying sentiment of a text message, i.e., a tweet. Twitter sentiment or opinion expressed through it may be positive, negative or neutral. However, no algorithm can give you 100% accuracy or prediction on sentiment analysis.

As a part of Natural Language Processing, algorithms like SVM, Naive Bayes is used in predicting the polarity of the sentence. sentiment analysis of Twitter data may also depend upon sentence level and document level.

Methods like, positive and negative words to find on the sentence is however inappropriate, because the flavor of the text block depends a lot on the context. This may be done by looking at the POS (Part of Speech) Tagging.

2.2.4.2Why Twitter Sentiment Analysis?

Sentiment Analysis Dataset Twitter has a number of applications:

A. Business: Companies use Twitter Sentiment Analysis to develop their business strategies, to assess customers' feelings towards products or brand, how people respond to their campaigns or product launches and also why consumers are not buying certain products.
B. Politics: In politics Sentiment Analysis Dataset Twitter is used to keep track of political views, to detect consistency and inconsistency between statements and actions at the government level. Sentiment Analysis Dataset Twitter is also used for analyzing election results.
C. Public Actions: Twitter Sentiment Analysis also is used for monitoring and analyzing social phenomena, for predicting potentially dangerous situations and determining the general mood of the blogosphere.

2.2.4.3Twitter Sentiment Analysis in R

R, a programming language intended for deep statistical analysis, is open source and available across different platforms, e.g., Windows, Mac, Linux. You can use R to extract and visualize Twitter data. You can create an app to extract data from Twitter.

Prerequisites for creating an app for extracting data for Twitter Sentiment Analysis in R

- R must be installed and you should be using RStudio.
- In order to extract tweets, you will need a Twitter application and hence a Twitter account. If you don't have a Twitter account, please sign up.
- Use your Twitter login ID and password to sign in at Twitter Developers.

Follow the steps below:

Once you have your twitter app setup, you are ready to dive into accessing tweets in R

You will use the retweet package to do this.

```
# load twitter library
Library(rtweet)
# plotting and pipes
Library(ggplot2)
Library(dplyr)
# text mining library
Library(tidytext)
```

The first thing that you need to set up in your code is your authentication. When you set up your app, it provides you with 3 unique identification elements:

1. appnam
2. key
3. secret

These keys are located in your twitter app settings in the Keys and Access Tokens tab. You will need to copy those into your code. Next, you need to pass a suite of keys to the API.

Finally, you can create a token that authenticates access to tweets! Note that the authentication process below will open a window in your browser.

2.2.4.4Tools for Twitter Sentiment Analysis:

2.2.4.4.1 Enginuity:

Enginuity, even though a paid solution, a basic version is available as a free web application. It works differently from many of the free sentiment analytics tools out there. Instead of directly querying tweets related to a certain keyword, Enginuity allows you to search for recent news stories about the keyword.

The tool then queries both Twitter and Facebook to calculate how many times the story has been shared. It also analyzes whether the sentiment of social shares is positive or negative, and gives an aggregate sentiment rating for the news story.

Enginuity is an awesome tool for finding stories to share through your social channels, as well as getting a combined picture of sentiment about recent events trending on social media.

2.2.4.4.2 Revealed Context (API/Excel Add-in):

Revealed Context, another popular tool for sentiment analytics on Twitter data, offers a free API for running sentiment analytics on up to 250 documents per day. There's an Excel add-in as well as a web interface for running analytics independently of the API.

While Revealed Context does not offer an interface for directly scraping Twitter, it can, however, analyze a spreadsheet of tweets without using the API. With the API, you can build a pipeline that feeds recent tweets from the Twitter API into the Revealed Context API for processing.

2.2.4.4.3 Steamcrab:

Steamcrab is a well-known web application for sentiment analytics on Twitter data. It focuses on keyword searches and analyzes tweets according to a two-pole scale (positive and negative). Visualization options are limited to scatter plots and pie charts.

2.2.4.4.4 MeaningCloud (API/Excel Add-in):

MeaningCloud is another free API for twitter text analytics, including sentiment analytics. One of the principal advantages of MeaningCloud is that the API supports a number of text analytics operations in addition to sentiment classification. These operations include topic extraction, text classification, part-of-speech tagging etc.

2.2.4.4.5 SocialMention (Web App):

Socialmention is a basic, search engine-style web app for topic-level sentiment analysis on Twitter data. You can enter a keyword, and the tool will return aggregate sentiment scores for the keyword as well as related keywords.

Another attractive feature of SocialMention is its support for basic brand management use case. It returns a "passion" score that measures how likely Twitter users are to discuss your brand, as well as the average reach of the Twitter users discussing your brand.

2.3 Related Work

A large number of studies have been involved in automatic prediction of diseases in social media content using machine learning approach.From the previous researches studies we mentioned some researches:

2.3.1Disease Detection

Brownstein et al. used online news, to perform surveillance of epidemics [25]. Their system, Healthmap, collects reports from online news aggregators, such as Google News. By categorizing the news into epidemics-related and unrelated reports, and filtering the epidemics-related documents into "breaking news,""warnings," and "old news," the system is able to trigger alerts based on the "breaking news."

Collier et al. developed a model [26] to automatically classify Twitter messages into six fixed syndromic classes, such as Respiratory and Gastrointestinal. Aramaki et al. [27] used different ML models to classify influenza-related tweets into two categories (positive or negative). Among the models they experimented, the SVM model achieved the highest F-Measure of 0.756 in distinguishing relevant tweets from tweets that are irrelevant. Signorini et al. used a SVM-based estimator to analyze H1N1-related tweets [28], and estimated the ILI rate prior to official announcement by one to two weeks. Similarly, Culotta experimented with a number of regression models to correlate Twitter messages with CDC statistics [29] and provided a relatively simple method to track the ILI rate using a large number of Twitter messges [30].

Lampos et al. [31] used an approach to automatically learn a set of markers to help - compute flu scores, and achieved a high correlation with HPA flu score, which is the equivalent of the CDC score in the UK.

2.3.2 Sentiment Classification in Twitter

In sentiment analysis, Pandey and Iyer [32] claimed the significance of domain specific features other than common text features used in traditional information retrieval tasks. Barbosa and Feng [33] focused on automation of the training data generation process.

Their work combined sentimentlabelled tweets coming from three sources: Twendz, Twitter Sentiment, and Tweet Feel. A moderate Cohen's kappa coefficient served as evidence that the combination of sources reduced the bias of the individual sources. In this way, the combination improved the polarity classification. The Naïve Bayes classifier is reported by Yu et al. [34] as the best interms of precision and recall, when applied to sentiment classification of news articles. Ardon et al. [35] researched the information spread in social networks and concluded that if a topic is supposed to be popular, it must cross regional borders aggressively. However, these studies do not analyze the health sentiments.

Salathé and Khandelwal [36] applied sentiment analysis to Twitter users' reaction towards the H1N1 vaccine. They classified Twitter messages into four categories: positive, neutral, negative, and irrelevant, and then calculated the H1N1 vaccine sentiment score from the relative difference of positive and negative messages. Chew and Eysenbach [37] used a statistical approach to computing the relative proportion of all tweets expressing concerns about H1N1 and visualized the temporal trend.

The baseline of study has been carried out on a dataset, which contains 5193 tweets from Twitter using API. In this study, following classifiers were investigated:
Support Vector Machine and Random Forest. After the experiments of classifier model SVM was found to be an efficient classifier for predicting the diseases names from the text of tweets.

Chapter Three: Research Methodology

3.1 Introduction

At this chapter, building the classifier model for prediction disease name from tweets based on text analysis of Twitter, description of the steps research methodology who followed to solve problem statement.

3.2 Methodology

There are no classifier models dedicated to the mining tweets, but there are some studies talking about Sentiment analysis. Therefore, the research idea concerned about a building classifier model for disease discovery from tweets.

3.2.1 Twitter Data Collection

Twitter is a fast growing microblogging website. Millions of people share their thoughts publicly. The short text message is known as a tweet, which is a max 140 characters in length. The hashtag is a common pound (#) symbol which is used to describe particular interest or group on social networking website. So, Twitter data were our main target for further analysis.

Extracted almost 5194 tweets over a period of two days. Twitter provides two API's for the extraction of data, REST and Streaming. Rest API is used to access previous tweets of users, whereas the Streaming API allows one to access tweets in real time based on a certain query. We used Streaming API to access tweets in real time manner based on hashtags, users etc. For this purpose, Twitter provides OAuth to access this APIs. First, user needs to create a Twitter application and then generate a consumer key, a consumer secret key, an access token and an access token secret key, which enable users to access the Twitter API on behalf of them (Twitter, 2015).

Classifier model was designed using R programming language. '**twitteR**' which is an open source package is used to extract tweets from Twitter. It allows R to use Twitter APIs to access tweets.

Following is a diagrammatic representation of data extraction process:

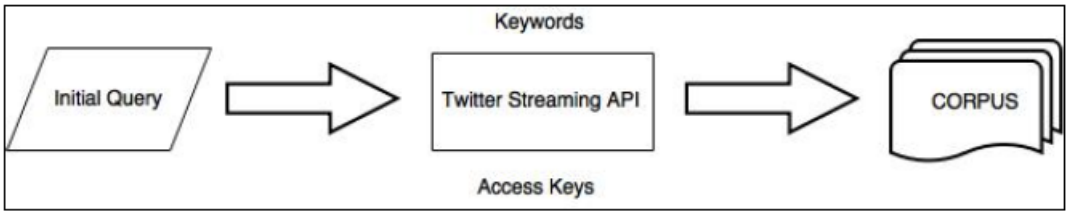


Figure (3.1): Data Extraction.

Initial Query is to start initial streaming process using Twitter API. Twitter streaming API initiates the process of data extraction using certain keywords and access keys. Corpus is nothing but tweets, which are stored in R format (Twitter, 2015).
We used around 7 different keywords to extract the data. Keywords are as follows: blood pressure, bone, cancer, diabetes, kidney, lung cancer and migraines. Based on these keywords, we segregated tweets into 7 different categories for further evaluation of tweets.

Categories are as follows: 0) blood pressure, 1) bone, 2) cancer, 3) diabetes, 4) kidney, 5) lung cancer and, 6) migraines.

Table 3.1. Sample Tweets Collected

No	Tweet	Topic
1	Health News FDA warns of serious genital infection linked to certain diabetes drugs FoxNews health wellness FDA https t co xGc2X MFF	0
2	mike hannay AHSN Network jameson I hope AHSNs see themselves as key bodies to help spread innovation diabetes	0
3	Great cause diabetes blakrozecomedy nccomedian https t co SqTjrq1nyW	0
4	I cut back on these because of my heartdisease and diabetes but I will mention that my skin has improved Somethi https t co KE7b9b4Jpl	0
5	RT parthaskar Mental Health Diabetes Come and contribute to NHS Diabetes Prog starting work on better support for those living with diabetes	0
6	RT European Cancer European Cancer IRQCC programme aims to improve outcomes for cancer patients in EU through the adoption and impleme	1
7	My favorite part meeting new kids who also have parents w cancer Cancer Pathways https t co 2xWpdiq246 https t co l0yf4zcoz4	1
8	My goal is to see that mental illness is treated like cancer Jane Pauley mentalhealth mentalillness https t co P7bp21Bywo	1
9	RT CancerTerms Thats why if a Cancer doesn't like you they make sure you dont touch them even accidentally they are very conscious of that	1

10	12 year old Ohio State fan Grant Reed nicknamed the cancer tumor in his brain Michigan then he beat it pun	1
11	Quitting Smoking Dramatically Reduces Diabetes Related Heart Disease Risk Study Suggests from EverydayHealth https t co yWScOrrc4x	2
12	Fats of Life Recent studies show that dietary fat is not the cause of rising cases of heart diseases Experts s https t co WSl2yYULkV	2
13	Are you a healthcare professional working in diabetes Test out your knowledge on the link between cardiovascular https t co MySvr6GZsZ	2
14	How does diabetes affects your body Infographic https t co 4c2ro28ixs https t co 05BRgdWR4B	2
15	For people with diabetes to be safe in hospital we need 1 Strong clinicalleadership 2 Access to an MDFT 3 HC https t co kgxKIjr7sa	2
16	RT Healing Soda affects the body in a chain reaction of health hazards Sugar Kills Diabetes Health Poison Diet pop https t co Yw	3
17	Study in Jan Feb 2018 issue of AnnFamMed https t co OfNkhwetNo examines primary care communication w patients https t co a6k7fLPOAl	3
18	Read New Insight Into Obstructive Sleep Apnea and Diabetic Retinopathy from the September diabetes issue by Drs https t co cXfx33v3mj	3
19	Bulletproof Workout Warmup Routine justhuynh fitness obesity health diabetes nutrition gym fatloss https t co pHLvo1lAro	3
20	RT anskaityte A WHO report estimates that more than a quarter of people worldwide 1 4 billion are not doing enough physical exercise	3
21	RT BloodDonorsIn Delhi Need Blood any group 6 units for Cancer fighter at AIIMS Hospital Call 7979810123 9084013536 via iRJPrateek	4
22	Cancer s have a troubled past which they reflect on over blunts	4

	and booze	
23	Deep learning detects segments classifies breast tumors with 93 accuracy https t co suGUpBT7uY DeepLearning Cancer Imaging	4
24	Cancer Zodiac People A Cancer woman can anger easily over the smallest of mishaps	4
25	Keeping some of the old and incorporating the new Cancer grief https t co OZLMwUmI6C	4
26	RT VABVOX I have cancer I am a feminist I can t think of two things more dissimilar Cancer takes lives every day after brutal terri	5
27	British Journal of Pharmacology Cannabis Effective Treatment For Cancer In Pre Clinical Studies Urgent need for https t co NoZyBbWrUJ	5
28	Has anyone taken HPV post 30yrs old If yes pls DM Need more info on the need benefits vaccination Cancer	5
29	RT Protectcare Even after cancer forced him to miss more than a year of high school one senior will be the first in his family to gradu	5
30	What signs of breastcancer indicate a need for surgery Your doctor will look out for symptoms such as skin dimpli https t co QQFYuI0rE9	5
31	Only able to sleep for 4 hours despite spending all day yesterday in tears amp being exhausted I have panic sitting https t co IbA6WY8709	6
32	Amazing Benefits of Pineapple for Cancer Arthritis and More https t co 36NGUmu9yc	6
33	Monsanto food poison CriminalMinds crimes justice fakefood cancer environment farmers earth humanity https t co Xj59q7e70m	6
34	via NatGeo DDT Pesticide Linked to Fourfold Increase in BreastCancer Risk https t co ZUNz0tsh7V Cancer https t co rrp6b5OqOG	6
35	The HPVvaccine jab that protects against a virus that causes cervical cancer will be given to boys aged 12 to https t co iLr8mjSH7O	6

In order to build a classifier model used to evaluate quality of classifier algorithms the following steps need to be followed see figure (3.2)

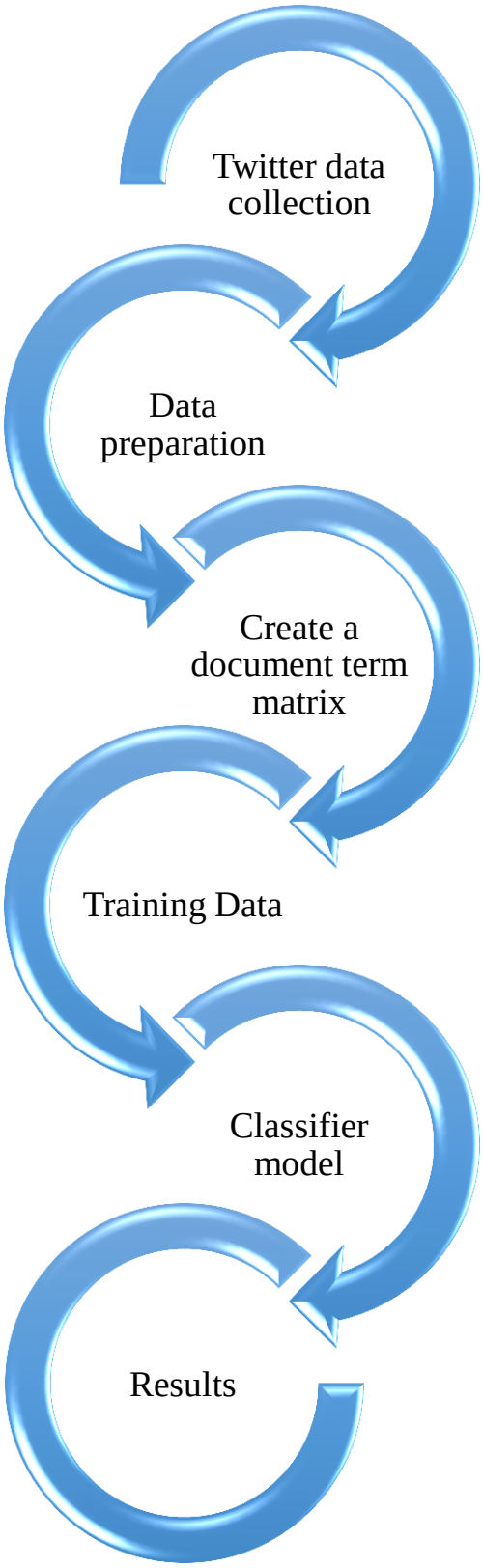


Figure (3.2) The Research Methodology.

3.2.2 Data Preparation

Extracted tweets were in R format. We filtered few objects from the R format such as Tweet Text, Created at (Date and Time), Location and Username.

We then removed hashtag # followed by RT@ and @ symbols. Regular expressions, special characters were also removed. We also removed http:// and the following web address from the text. We also remove punctuation, stop words and numbers. All the tweets were then turned to lowercase letters.
Specifically, we replaced following symbols/text with blank spaces:

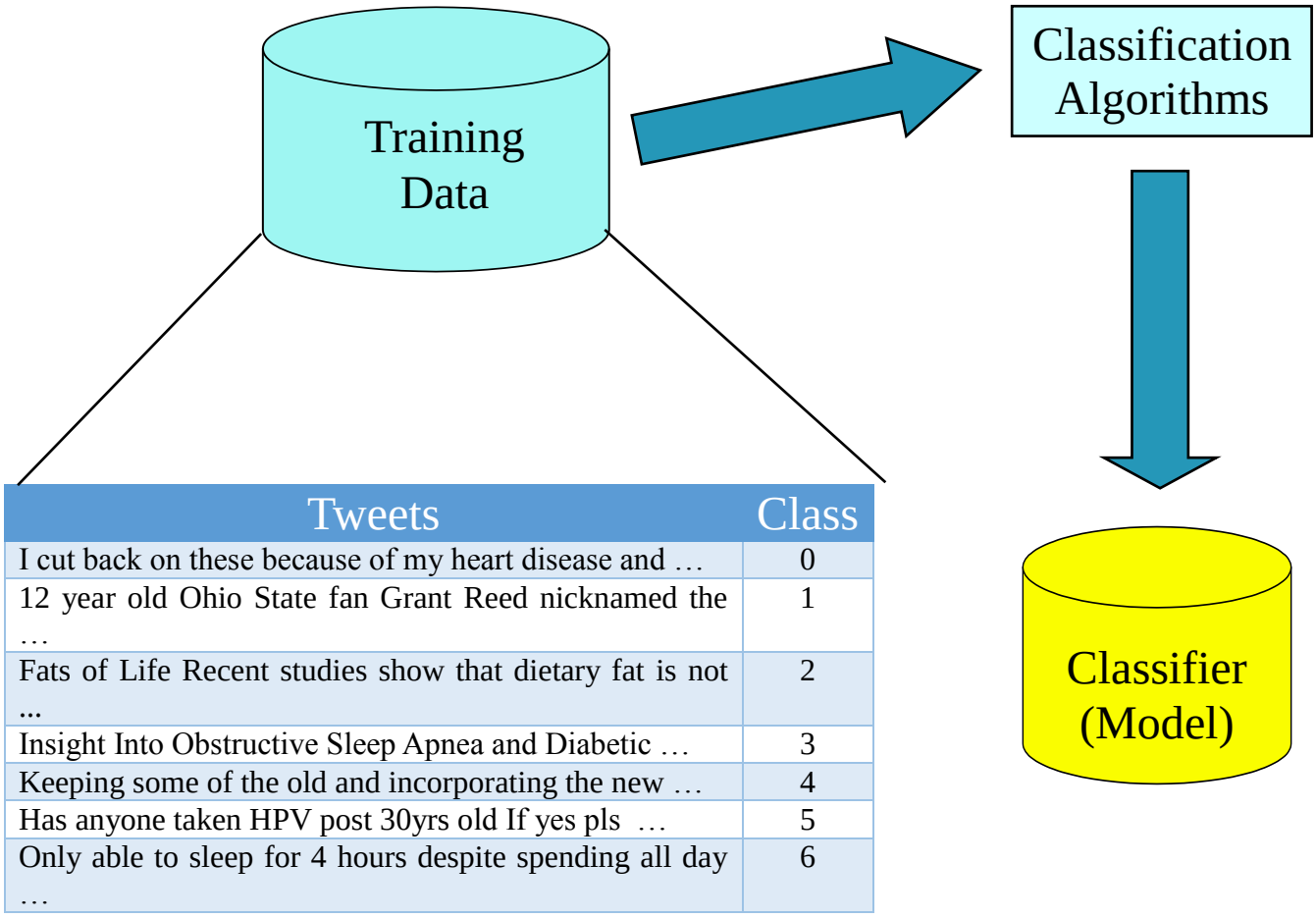
1. Remove text which starts from http[^\\s]+
2. Remove hashtag #, @, RT@
3. Remove Unicode characters /([\ud800-\udbff][\udc00-\udfff])|./g
4. Convert to lowercase
5. Remove hyperlinks
6. Remove punctuation
7. Remove stop words
8. Remove numbers

We did not remove Retweets, since it could be very useful in our analysis. Retweet is, when the user wants to repost someone else's tweet, she/he simply retweets. So, a user might feel the same thing as the other person. This could play an important contribution towards our study. Pre-processing and cleaning of the text also plays very significant role, since it removes all the inconsistency and irrelevant data from the text, which could lead to false results.

3.2.3 Create a Document Term Matrix

In text mining, it is important to create the document-term matrix (DTM) of the corpus we are interested in. A DTM is basically a matrix, with documents designated by rows and words by columns, that the elements are the counts or the weights (usually by tf-idf). Subsequent analysis is usually based creatively on DTM.

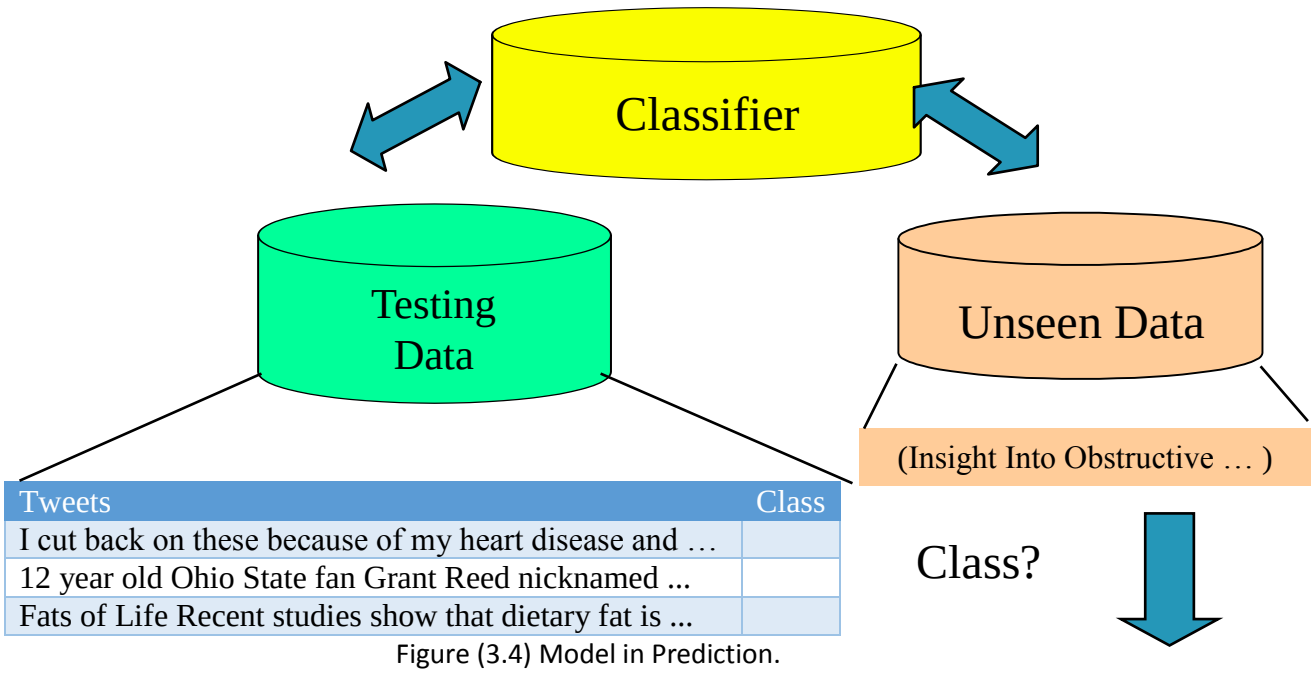
3.2.4 Training Data



Tweets	Class
I cut back on these because of my heart disease and …	0
12 year old Ohio State fan Grant Reed nicknamed the …	1
Fats of Life Recent studies show that dietary fat is not ...	2
Insight Into Obstructive Sleep Apnea and Diabetic …	3
Keeping some of the old and incorporating the new …	4
Has anyone taken HPV post 30yrs old If yes pls …	5
Only able to sleep for 4 hours despite spending all day …	6

Figure (3.3) Model Construction.

3.2.5 Classifier Model



Tweets	Class
I cut back on these because of my heart disease and …	
12 year old Ohio State fan Grant Reed nicknamed ...	
Fats of Life Recent studies show that dietary fat is ...	

Figure (3.4) Model in Prediction.

3

Chapter Four: Results and Discussions

4.1 Introduction

In this chapter, we discuss the results of classifier models at the point of prediction of disease name by used two algorithms Support Vector Machine and Random Forest.

4.2 Evaluation Matrices

In this section, define the evaluation metrics used in each algorithms of classification.

4.2.1 Confusion Matrix

A confusion matrix is the measure of performance of a multi-label/multi-class classification model. It is also referred to as error matrix or a table of confusion. This is used in predictive analytics to understand what type of data is being labeled as 'true' and what kind of data is labelled as 'false' by the classifier or the classification model chosen. In summary, Confusion matrix tells us how the classification algorithm is performing with respect to the ground.
This matrix reports the number of true positives, false positives, false negatives, and true negatives from the classifier. The terminology used in the confusion matrix is defined as follows:

- **True Positive (TP)**: The number of instances which are correctly labelled as positive.
- **False Negative (FN):** The number of instances which are predicted as negative but in reality these instances are positive.
- **False Positive (FP):** The number of instances which are predicted as positive but in reality these instances are negative.
- **True Negative (TN):** The number of instances which are correctly labelled as negative.

A pictorial representation of confusion matrix is shown in Table 4.1.

		Predicted Class	
		Predicted True	**Predicted False**
Actual Class	**Actual True**	True Positive	False Negative
	Actual False	False Positive	True Negative

Table 4.1: Confusion matrix representation

4.2.2 First Stage Classification Evaluation

- **Precision:** In case of binary classification, the precision of the classifier is defined as the ratio of number of true positives over the number of false positives plus the number of true positives.

  Precision (Pbinary) = TPTP+FP

- **Recall:** Recall of the classifier is defined as the ratio of the number of true positives over the number of false negatives plus true positives.

  Recall (Rbinary) = TPTP+FN

- **F-measure:** F-measure of the classifier is defined as the harmonic mean of precision and recall. The formula for calculating F-measure is same for first and second stage classification.

  F- measure (F) = 2PRP+R

4.2.3 Second Stage Classification Evaluation

In this section, define the evaluation criteria used for multi-label classification which used in second-stage of classification model [25].
n= number of instances in the dataset
Yi = actual label of the ith instance
Zi = predicted label of the ith instance

- **Precision:** In case of multi-label classification, the precision is defined as the proportion of predicted correct labels to the total number of actual labels, averaged over all instances.

  Precision (Pmulti-label) = 1ni=1n|Yi∩Zi||Zi|

- **Recall:** Recall is defined as the proportion of predicted correct labels to the total number of predicted labels, averaged over all instances.

  Recall (Rmulti-label) = 1ni=1n|Yi∩Zi||Yi|

- **Exact Match:** Exact match is defined as the average of the count of sample where predicted label equals to actual labels.

  Exact Match (EMmulti-label) = 1ni=1nI(Yi =Zi)
  Where **I** is the indicator function, which gives a value 1 if Yi = Zi and a value 0 if Yi Zi.

- **Accuracy:** Accuracy for each instance is defined as the proportion of the predicted correct labels to the total number (predicted and actual) of labels for that instance. Overall accuracy is the average across all instances.

  Accuracy (Amulti-label) = 1ni=1n|Yi∩Zi||Yi∪ Zi|

4.3 Classification Algorithms

4.3.1 Support Vector Machine

Support Vector Machines are an excellent tool for classification, novelty detection, and regression. ksvm supports the well-known C-svc, nu-svc, (classification) one-class-svc (novelty) eps-svr, nu-svr (regression) formulations along with native multi-class classification formulations and the bound-constraint SVM formulations. ksvm also supports class-probabilities output and confidence intervals for regression.

4.3.1.1 Modelling Result

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1   2   3   4   5   6
         0 180   0  57  52   0   0   0
         1   0 155   0   0  26  39  33
         2   8   0 129  30   0   0   0
         3  12   0  14 118   0   0   0
         4   0  19   0   0 138  19  20
         5   0  19   0   0  28 139  32
         6   0   7   0   0   8   3 115

Overall Statistics

               Accuracy : 0.6957
                 95% CI : (0.6709, 0.7197)
    No Information Rate : 0.1429
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.645
 Mcnemar's Test P-Value : NA
```

Table 4.1: Statistics by Class:

	Class: 0	Class: 1	Class: 2	Class: 3	Class: 4	Class: 5	Class: 6
Sensitivity	0.9000	0.7750	0.6450	0.59000	0.69000	0.69500	0.57500
Specificity	0.9092	0.9183	0.9683	0.97833	0.95167	0.93417	0.98500
Pos Pred Value	0.6228	0.6126	0.7725	0.81944	0.70408	0.63761	0.86466
Neg Pred Value	0.9820	0.9608	0.9424	0.93471	0.94850	0.94839	0.93291
Prevalence	0.1429	0.1429	0.1429	0.14286	0.14290	0.14286	0.14286
Detection Rate	0.1286	0.1107	0.0921	0.08429	0.09857	0.09929	0.08214
Detection Prevalence	0.2064	0.1807	0.1193	0.10286	0.14000	0.15571	0.09500
Balanced Accuracy	0.9046	0.8467	0.8067	0.78417	0.78417	0.81458	0.78000

4.3.2 Random Forest Method of Classification

Random Forest is considered a machine learning algorithm that generates a set of decision trees using sampling with replacement in the original dataset. The number of decision trees generated is specified by the programmer. This algorithm gives rise to a model that results from the voting among all the generated random trees. This means that, for example, a certain attribute "a" belongs to class "1", if this attribute was predicted by most trees generated as belonging to class"1".

4.3.2.1 Modelling Result

Call:
randomForest(formula = as.factor(corpus) ~ ., data = df.train, importance = TRUE)

Type of random forest: classification Number of trees: 500

No. of variables tried at each split: 68 OOB estimate of error rate: 79.5%

Table 4.2: Confusion matrix:

	0	1	2	3	4	5	6	class.error
0	41	1	84	73	0	1	0	0.795
1	4	58	4	7	34	41	52	0.710
2	92	0	35	71	0	0	2	0.825
3	94	1	75	29	0	1	0	0.855
4	4	50	3	5	42	39	57	0.790
5	4	52	3	6	37	38	60	0.810
6	5	51	6	10	39	45	44	0.780

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1   2   3   4   5   6
         0 160   0  42  40   0   0   0
         1   0 150   0   0  20  29  22
         2  21   0 140  32   0   0   0
         3  19   0  18 128   0   0   0
         4   0  19   0   0 145  23  23
         5   0  12   0   0  17 124  17
         6   0  19   0   0  18  24 138

Overall Statistics

               Accuracy : 0.7036
                 95% CI : (0.6789, 0.7274)
    No Information Rate : 0.1429
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6542
 Mcnemar's Test P-Value : NA
```

Table 4.3: Statistics by Class:

	Class: 0	Class: 1	Class: 2	Class: 3	Class: 4	Class: 5	Class: 6
Sensitivity	0.8000	0.7500	0.7000	0.64000	0.72500	0.62000	0.69000
Specificity	0.9317	0.9408	0.9558	0.96917	0.94580	0.96167	0.94917
Pos Pred Value	0.6612	0.6787	0.7254	0.77576	0.69050	0.72941	0.69347
Neg Pred Value	0.9655	0.9576	0.9503	0.94170	0.95380	0.93821	0.94838
Prevalence	0.1429	0.1429	0.1429	0.14286	0.14290	0.14286	0.14286
Detection Rate	0.1143	0.1071	0.1000	0.09143	0.10360	0.08857	0.09857
Detection Prevalence	0.1729	0.1579	0.1379	0.11786	0.15000	0.12143	0.14214
Balanced Accuracy	0.8658	0.8454	0.8279	0.80458	0.83540	0.79083	0.81958

Chapter Five: Conclusions and Recommendations

5.1 Conclusions

As evidenced by the results provided in the study, diseases can be predicted its by analysing the posts on social networking sites.
As evidenced by the results provided in the study yes we can predication disease name through text mining for tweets to prediction disease name based on text analysis.

In this thesis, we built a classification model for disease name discovery from tweets. We collected the tweets using Twitter API. We created a balanced dataset for classification purposes. We developed a two-stage classification model discerning whether a tweet is disease name related or not. We showed significant improvement on Confusion Matrix and Overall Statistics. We used two classification algorithms Support Vector Manchin (SVM) and Random Forest. We found Random Forest was better than SVM in the accuracy of classifier model.

5.2 Recommendations

As complemented to this study, there are some recommendations for researcher in this subject to improve accuracy model:

- Future work Try to extract data form Arabic Text.
- Increasing dataset to help model to prediction by best accuracy.
- Using many method to split dataset for training and testing.

6.Reference

1. 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-

2. "Twitter." Wikipedia. http://en.wikipedia.org/wiki/Twitter. (accessed October 14, 2014).

3. B. A. Huberman, D. M. Romero, and F. Wu. "Social networks that matter: Twitter under the microscope." *arXiv preprint arXiv:0812.1045*, 2008.

4. "Hashtag Twitter." Twitter Help Center, 2014. https://support.twitter.com/articles/49309-using-hashtags-on-twitter. (accessed October 14, 2014).

5. H. Kwak, C. Lee, H. Park, and S. Moon. "What is Twitter, a social network or a news media?" Proceedings of the 19th international conference on World wide web, pp. 591-600, 2010.

6. C. Smith. "By The Numbers: 215 Amazing Twitter Statistics." http://expandedramblings.com/index.php/march-2013-by-the-numbers-afew-amazing-twitter-stats/ - .U-5fWLxdV_E.

7. "Twitter Limits." Twitter Help Center. https://support.twitter.com/articles/15364-twitter-limits-api-updates-andfollowing. (accessed October 11, 2014).

8. "Representational state transfer." Wikipedia. http://en.wikipedia.org/wiki/Representational_state_transfer. (accessed October 12, 2014).

9. "The Twitaholic.com Top 100 Twitterholics based on followers." Twitaholic. http://twitaholic.com/. (accessed October 12, 2014).

10. "The Klout score." Klout. https://klout.com/corp/score. (accessed October 12, 2014).

11. E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts. "Everyone's an influencer: quantifying influence on twitter." *Proceedings of the fourth ACM international conference on Web search and data mining*, pp. 65-74,2011.

12. M. Trusov, A. V. Bodapati, and R. E. Bucklin. "Determining influential users in internet social networks." *Journal of Marketing Research*, vol. 47, no. 4, pp. 643-658, 2010.

13. "The Emerging Science of Superspreaders (And How to Tell If You're One Of Them)." MIT Technology Review, 2014. http://www.technologyreview.com/view/527271/the-emerging-science-ofsuperspreaders-and-how-to-tell-if-youre-one-of-them/.

14. I. Anger, and C. Kittl. "Measuring influence on Twitter." Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies. ACM, pp. 31, 2011.

15. A. Neuhauser. "Health Care Harnesses Social Media." U.S.NEWS, 2014. http://www.usnews.com/news/articles/2014/06/05/health-care-harnessessocial-media.

16. C. Hawn. "Take two aspirin and tweet me in the morning: how Twitter, Facebook, and other social media are reshaping health care." *Health affairs*,vol. 28, no. 2, pp. 361-368, 2009.

17. P. H. Keckley. "Social Networks in Health Care, Communication, collabortation and insights." Deloitte Center for Health Solutions, 2010. http://www.ucsf.edu/sites/default/files/legacy_files/US_CHS_2010SocialNetworks_070710.pdf. (accessed October 20, 2014).

18. E. Quincey, and P. Kostkova. "Early warning and outbreak detection using social networking websites: The potential of twitter." *Electronic healthcare*, pp. 21-24, 2010.

19. M. J. Paul, and M. Dredze. "You are what you Tweet: Analyzing Twitter for public health." *ICWSM*, pp. 265-272, 2011.

20. J. Lin, and A. Kolcz. "Large-scale machine learning at twitter." *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*. ACM, pp. 793-804, 2012.

21. D. Alexander. "Data Mining." http://www.laits.utexas.edu/~anorman/BUS.FOR/course.mat/Alex/ - 3. (accessed October 19, 2014).

22. U. Fayyad, and R. Uthurusamy. "Evolving data into mining solutions for insights." *Communications of the ACM*, vol. 45, no. 8, pp. 28-31, 2002.

23. K. Kranz. "Dig'in Social Media-The Data Mining of Social Media with Hadoop and his Friend Mahout." 2013. http://www.ca.com/us/~/media/Files/About%20Us/CATX/digin-socialmedia-kranz.pdf.

24. https://www.digitalvidya.com/blog/twitter-sentiment-analysis-introduction-and-techniques/

25. Spitzer, Robert L., et al. "Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study." *Jama*282.18 (1999): 1737-1744.

26. Ford, Daniel E., and Thomas P. Erlinger. "Depression and C-reactive protein in US adults: data from the Third National Health and Nutrition Examination Survey."

27. https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

28. Hosmer Jr, David W., Stanley Lemeshow, and Rodney X. Sturdivant. *Applied logistic regression*. Vol. 398. John Wiley & Sons, 2013.

29. US Department of Health and Human Services. "National Institute of Mental Health.(2011). Depression (NIH Publication No. 11-3561)."

30. https://wordnet.princeton.edu/documentation/morphy7wn

31. Yazdavar, Amir Hossein, et al. "Semi-Supervised Approach to Monitoring Clinical Depressive Symptoms in Social Media." *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*. ACM, 2017.

32. V. Pandey and C.V.K. Iyer, "Sentiment Analysis of Microblogs," Technical Report, Stanford University, 2009.

33. L. Barbosa, and J. Feng, "Robust Sentiment Detection on Twitter from Biased and Noisy Data," In Proceedings of the 23rd International Conference on Computational Linguistics:Posters, 2010.

34. H. Yu and V. Hatzivassiloglou, "Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences," In Proceedings of the 2003 conference on Empirical methods in natural language processing, 2003.

35. S. Ardon, A. Bagchi, A. Mahanti, A. Ruhela, A. Seth, R. Tripathy, and S. Triukose, "Spatio-Temporal Analysis of Topic Popularity in Twitter," arXiv:1111.2904v1 [cs.SI], 2011.

36. M. Salathé and S. Khandelwal, "Assessing Vaccination Sentiments with Online Social Media: Implications for Infectious Disease Dynamics and Control," PLoS Comput Biol 7(10): e1002199. doi:10.1371/journal.pcbi. 1002199,2011.

37. Chew and G. Eysenbach, "Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1Outbreak," PLoS ONE 5(11): e14118. doi:10.1371/journal. pone.0014118, 2010.

7.Appendices

APPENDIX A

R CODE TO DATA HANDLING

7.1 R CODE TO DATA HANDLING

7.1.1 LOADING TWITTER DATA:

```
load('data/blood pressure_data.RData')
blood_pressure<-d

load('data/bone diseases.RData')
bone<-d

load('data/cancer_data.RData')
cancer<-d

load('data/diabetes_data.RData')
diabetes<-d

load('data/kidney.RData')
kidney<-d

load('data/lungcancer.RData')
lungcancer<-d

load('data/Migraines.RData')
Migraines<-d

rm(d)
```

7.1.2 Merging Twitter Data:

```
diseases<-cbind.data.frame(blood_pressure$text                                    ,
rep(0,length(blood_pressure$text)))
colnames(diseases)<-c('twitte', 'topic')

temp<-cbind.data.frame(bone$text , rep(1,length(bone$text)))
colnames(temp)<-c('twitte', 'topic')

diseases<-rbind.data.frame(diseases,temp)

temp<-cbind.data.frame(cancer$text , rep(2,length(cancer$text)))
colnames(temp)<-c('twitte', 'topic')

diseases<-rbind.data.frame(diseases,temp)

temp<-cbind.data.frame(diabetes$text , rep(3,length(diabetes$text)))
colnames(temp)<-c('twitte', 'topic')

diseases<-rbind.data.frame(diseases,temp)

temp<-cbind.data.frame(kidney$text , rep(4,length(kidney$text)))
colnames(temp)<-c('twitte', 'topic')

diseases<-rbind.data.frame(diseases,temp)

temp<-cbind.data.frame(lungcancer$text , rep(5,length(lungcancer$text)))
colnames(temp)<-c('twitte', 'topic')
```

```
diseases<-rbind.data.frame(diseases,temp)
temp<-cbind.data.frame(Migraines$text , rep(6,length(Migraines$text)))
colnames(temp)<-c('twitte', 'topic')

diseases<-rbind.data.frame(diseases,temp)

diseases$twitte<-as.character(diseases$twitte)
Encoding(diseases$twitte)  <- "UTF-8"


save(diseases,file = 'disease.RData')
```

7.1.3 Data Preparation

```
library("stringr", lib.loc="C:/Program Files/R/R-3.5.0/library")

x <- "a1~!@#$%^&*(){}_+:\"<>?,./;'[]-="
str_replace_all(x, "[[:punct:]]", " ")
diseases$twitte <- str_replace_all(diseases$twitte, "[[:punct:]]", " ")
diseases2 <- diseases[which(!grepl("[^\x01-\x7F]+", diseases$twitte)),]

table(diseases2$topic)
sample_index = sample(which(diseases2$topic == 0), 742)
sample_index = c(sample_index,sample(which(diseases2$topic == 1), 742))
sample_index = c(sample_index,sample(which(diseases2$topic == 2), 742))
sample_index = c(sample_index,sample(which(diseases2$topic == 3), 742))
sample_index = c(sample_index,sample(which(diseases2$topic == 4), 742))
sample_index = c(sample_index,sample(which(diseases2$topic == 5), 742))
sample_index = c(sample_index,sample(which(diseases2$topic == 6), 742))

diseases2<-diseases2[sample_index,]

table(diseases2$topic)

save(diseases2,file = 'disease_sampled.RData')

corpus <- VCorpus(VectorSource(diseases2$twitte))
```

7.1.4 Create a Document Term Matrix

```
corpus <- tm_map(corpus, content_transformer(removePunctuation))

corpus <- tm_map(corpus, content_transformer(removeWords),
                 stopwords("english"))

corpus <- tm_map(corpus, content_transformer(tolower))

corpus <- tm_map(corpus, content_transformer(removeWords),
                 stopwords("english"))

corpus <- tm_map(corpus, stemDocument)

corpus <- tm_map(corpus, stripWhitespace)

corpus <- tm_map(corpus, content_transformer(removeNumbers))

tdm <- DocumentTermMatrix(corpus)
table(diseases2$topic)
sample_index = sample(which(diseases2$topic == 0), 250)
sample_index = c(sample_index,sample(which(diseases2$topic == 1), 250))
sample_index = c(sample_index,sample(which(diseases2$topic == 2), 250))
sample_index = c(sample_index,sample(which(diseases2$topic == 3), 250))
sample_index = c(sample_index,sample(which(diseases2$topic == 4), 250))
sample_index = c(sample_index,sample(which(diseases2$topic == 5), 250))
sample_index = c(sample_index,sample(which(diseases2$topic == 6), 250))
samdiseases2<-diseases2[sample_index,]
table(samdiseases2$topic)
```

APPENDIX B

R CODE TO DATA MODELLING

7.2 R Code to Data Modelling

7.2.1 Import Libraries

```
library("kernlab")
library("caret")
library("tm")
library("dplyr")
library("splitstackshape")
library("e1071")
library(randomForest)
```

7.2.2 Training Data

```
train<-VCorpus(VectorSource(samdiseases2$twitte))#VCorpus(DirSource
("Training", encoding = "UTF-8"), readerControl=list(language="English"))
train <- tm_map(train, content_transformer(stripWhitespace))
train <- tm_map(train, content_transformer(tolower))
train <- tm_map(train, content_transformer(removeNumbers))
train <- tm_map(train, content_transformer(removePunctuation))

train.dtm<-as.matrix(DocumentTermMatrix(train,
control=list(wordLengths=c(1,Inf))))
train.df <- data.frame(train.dtm)
label.df <- data.frame(samdiseases2$topic)#data.frame(row.names(train.df))
colnames(label.df) <- c("filenames")
label.df<- cSplit(label.df, 'filenames', sep="_", type.convert=FALSE)
train.df$corpus<- label.df$filenames_1
df.train <- train.df
df.test <- train.df
```

7.2.3 Create a Support Vector Machine Model

```
df.model<-ksvm(corpus~., data= df.train, kernel="rbfdot")

df.pred<-predict(df.model, df.test)
```

7.2.3.1 Create a Confusion Matrix

```
con.matrix<-confusionMatrix(df.pred, as.factor(df.test$corpus))

print(con.matrix)
```

7.2.4 Create a Random Forest Model with Default Parameters

```
model1 <- randomForest(as.factor(corpus) ~ ., data = df.train, importance = TRUE)
model1

df.pred<-predict(model1, df.test)
```

7.2.4.1 Create a Confusion Matrix

```
con.matrix<-confusionMatrix(df.pred, as.factor(df.test$corpus))
print(con.matrix)
```