COMPUTATIONAL CONTENT ANALYSIS AND STUDY OF ZIKA VIRUS OUTBREAKS ON TWITTER

A project report submitted in partial fulfillment of the requirements for B.Tech. Project

B.Tech.

by

Arabh Kumar (2014IPG-020)
Vipin Kumar (2014IPG-103)
Buddh Priy Maury (2014IPG-116)

विश्वजीवनामृतं ज्ञानम्

ABV INDIAN INSTITUTE OF INFORMATION TECHNOLOGY AND MANAGEMENT

GWALIOR-474 010

2017

CANDIDATES DECLARATION

We hereby certify that the work, which is being presented in the report, entitled **Computational content analysis and study of zika virus outbreaks on twitter** , in partial fulfillment of the requirement for the award of the Degree of **Bachelor of Technology** and submitted to the institution is an authentic record of our own work carried out during the period *May 2017* to *September 2017* under the supervision of **Dr.Pradip Swarnakar**. We also cited the reference about the text(s)/figure(s)/table(s) from where they have been taken.

Date: Signatures of the Candidates

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

Date: Signatures of the Research Supervisors

ABSTRACT

In recent years social media like Twitter has become one of the major sources for opinion and information sharing, and this became popular with the affordable mobile devices and its portability with social media applications. People on social media like Twitter share their view on many topics, and this information is mined for various applications and predictions. One such application is real time disease surveillance. In this project, we have done real time disease surveillance regarding zika virus on Twitter. We have used Twitter API and various Python libraries to collect tweets on Zika virus from specific geographical location using a longitude, latitude value as center and radius value in order to cover desired area. We have used various Python libraries like Natural Language Toolkit (NLTK) and Naive Bayes Algorithm in order to find polarity of tweets and hence finding out level of concern among people from particular location (USA).We have analyzed the tweets based on six phrases in order to decide awareness level among people regarding Zika virus. Finally, we have found out based on percentage of negative tweets that now people are less concerned about Zika virus as expected.

Keywords: - Twitter, Sentiment Analysis, Zika Virus, Disease surveillance.

ACKNOWLEDGEMENTS

We are highly indebted to **Dr.Pradip Swarnakar**, and are obliged for giving us the autonomy of functioning and experimenting with ideas. We would like to take this opportunity to express our profound gratitude to them not only for their academic guidance but also for their personal interest in our project and constant support coupled with confidence boosting and motivating sessions which proved very fruitful and were instrumental in infusing self-assurance and trust within us. The nurturing and blossoming of the present work is mainly due to their valuable guidance, suggestions, astute judgment, constructive criticism and an eye for perfection. Our mentor always answered myriad of our doubts with smiling graciousness and prodigious patience, never letting us feel that we are novices by always lending an ear to our views, appreciating and improving them and by giving us a free hand in our project. It's only because of their overwhelming interest and helpful attitude, the present work has attained the stage it has.

Finally, we are grateful to our Institution and colleagues whose constant encouragement served to renew our spirit, refocus our attention and energy and helped us in carrying out this work.

(Arabh Kumar) (Vipin Kumar) (Buddh Priy Maury)

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION AND LITERATURE SURVEY

1.1 Introduction

In this chapter, we are going to focus on introductions on various aspects of our thesis.We are going to explain about Sentiment Analysis, Twitter Sentiment Analysis, various tools of Python which are useful for Sentiment Analysis and Natural Language Toolkit (NLTK). Then we will describe the objective of our thesis.

1.1.1 Sentiment analysis

Sentiment analysis is that the domain of study of examine peoples thoughts, sentiments, evaluations, mindset, and feeling from written language. Sentiment analysis systems are utilized in almost each content as a result of thoughts are vital to the majority human activities. They're key influencers of our behaviors. Sentiment analysis uses tongue process and text analysis to spot and extract data from a few specific area of interest. Attributable to massive use of the social media such as blogs and social networking sites like Twitter the interest in sentiment analysis has raised to the next extent. There are several problems in Sentiment analysis. The first is that an opinion word that's thought-about to be positive in one situation could also be taken negatively in another scenario. The second challenge is that folks don't continually reveal their opinions in the same method.

1.1.2 Twitter Sentiment analysis

Twitter Sentiment Analysis is the process of determining whether a tweet is positive, negative or neutral [5]. It can be used to identify people's opinion towards a brand or public action through the use of variables such as context, tone, emotion, etc. A

researcher can use sentiment analysis to find out public opinion on any epidemic or health related issues. The health department can also use this analysis to gather critical knowledge of awareness level in public with respect to any particular epidemic (Zika Virus in this case).

1.1.3 Python

We have used programming language Python for our thesis which is high level and dynamic programming language. Python libraries have grown significantly in last ten years and some of which are specifically designed for data analysis [8]. We have used Python 3.6 version. There are various open source libraries are available which is compatible with Python 3.6.

Python is a simple programming language, but its simplicity does not limit its versatility. There are other programming languages for data analysis such as 'R' and 'MATLAB', but they are not as flexible as python.

1.1.4 Natural Language Toolkit (NLTK)

Natural Language Toolkit (NLTK) is a Python library which is distributed under the GPL open source license and it has been rewritten multiple times in order to take advantage of recent development in Python language [2]. NLTK is a group of multiple python scripts which is used for data classification, text processing, and tokenization. This library plays a major role to obtain sentiment from text data.

There are many functions which are very useful for data pre-processing.These functions are used to preprocess the twitter data to make them fit for extracting features. NLTK works well with various machine learning algorithms which are used for Sentiment classification.

We used Python as the main programming language in which we have written our script for fetching tweets as well as sentiment analysis. NLTK is the library which does the most important task to classify text into either positive or negative or neutral class.

1.1.5 Naive Bayes Classifier

The task of supervised machine learning is to infer a function from tagged training samples. We are going to use Naive Bayes classifier which is a classifier with a probabilistic output. In such type of classifiers with probabilistic output we have option to reject with a probability distribution if we are not sure about prediction result and hence we can pass it for manual check up [10].

Suppose , x_1 to x_n is a dependent vectors and there is a class variable y. So according

to Bayes' probability theorem:

$$P(y|x_1, ..., x_n) = \frac{P(y)P(x_1, ..., x_n|y)}{x_1, ..., x_n} \tag{1.1}$$

According to independence assumption :

$$P(x_i|y, x_1, ..., x_(i-1), x_(i+1), ..., x_n) = P(x_i|y), \tag{1.2}$$

For each 'i' , this becomes

$$P(y|x_1, ..., x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i|y)}{P(x_1, ..., x_n)}. \tag{1.3}$$

Since P(x_1,.....,x_n) is constant on provided input , so we have classification rule as: P(y| x_1,.....,x_n)P(y) $\propto \prod_{i=1}^{n}$ P(x_i|y).

$$\hat{y} = arg\ max_y P(y)\prod_{i=1}^{n} P(x_i|y), \tag{1.4}$$

To evaluate we can use MAP (Maximum A Posterior) estimation P(y) and P(x_i |y) ; the P(y) of class 'y' is relative frequency in sample [20].

1.1.6 Zika Virus

Zika virus comes from virus family 'Flaviviridae' which is spread by Aedes mosquitoes during day time [7]. Zika virus is named after Zika forest situated in Uganda, the place where the virus was first isolated in 1947 [17]. Initially, no major symptoms are identified in Zika fever affected person [7]. Zika virus can also be spread through various other modes such as from mother to her infant, through sexual contact and through blood transfer from a Zika affected person [6]. There is no vaccine available for Zika virus [15].

1.2 Literature Review

In last few decades, multiple research has been performed on Sentiment analysis. Recently sentiment analysis on social media data has become very popular. Generally, sentiment analysis is done with the help of machine learning algorithms thus to find out whether a given text is positive or negative or neutral.

The first work on sentiment analysis aimed at classifying text by overall sentiment, not just focused on any one topic. They used various machine learning algorithms such as Naive Bayes, maximum entropy, and support vector machines (SVM). They found out that classification of sentiment is very tedious.They concluded that supervised machine learning algorithms are the foundation for sentiment analysis [12].

Collection of a large amount of data has been helpful to find out what people are thinking or presuming. Recently with the boom in social media sites, data available for opinion mining is very large. There are other recourses such as blogs, public comments on different sites etc. which are helpful to decide people's opinion about the topic. Various new systems are built depending on different coding languages as the work in the field of data mining is booming. Nowadays there are libraries and commands available which can perform live research [13].

Natural Language Toolkit (NLTK) is a library. This library is a combination of many script modules, a big set of structured files, different tutorials, numerous statistical functions, machine learning classifiers, etc. Natural Language Processing is the primary purpose of NLTK. Developers develop various new components and substitute them with existing component, more structured scripts are written, and better results are given by dataset [3].

Researchers performed a real time analysis of public responses for 2012 presidential election in U.S. to predict the election result. Their sentiment analysis was very fast compared to traditional content analysis. The system they explained is very effective for media and researchers [19].

French Polynesia went through the biggest Zika virus outbreak between 2013 and 2014. Increase in Guillain-Barre syndrome was identified during the period of the Zika virus outbreak. There was an expected relation between Zika virus and Guillain-Barre syndrome [4].

In traditional survey based methods, there is a big time gap but in new techniques of big social data mining help us to get rid of that time gap and also take care of privacy concerns in order to study public behavior on specific issues. In the past sentiment analysis has been performed for getting public views on many social issues such as gender based violence [16] and to find out health related opinions [14, 1].

1.2.1 Motivation

Twitter is a large social media channel where users tweet about various topic which also includes health issues. Traditional disease surveillance was done manually by selecting some target population and collecting their view about any particular disease. Social media channels, like Twitter, provides continuous information on public opinion about any epidemic and other health issues which can help public health agencies in performing real time surveillance.

1.2.2 Gap Analysis

There are some limitations which we noted during data set collection from twitter and analysis of that data set.

1. Since our script for data collection from Twitter does not understand and filter out sarcasm, the dataset of tweets also includes sarcastic tweets, and it is not possible for current sentiment analysis algorithms to accurately classify sarcasm into sentiment polarity.
2. Since our study is limited to very less number of languages, so it certainly limits the accuracy of our results.
3.In collected dataset, we noted that there were many tweets which do not make any sense in case of our study however it contains the keyword which we used to fetch the tweets from twitter.

1.3 Objective

Research Objective 1 - To examine level of concern on 'Zika virus' by analyzing sentiment polarity of tweets .
Research Question 1 - To what extent shared contents on twitter give legitimate information ?
Research Objective 2 - To find out what number of people are twitting about prevention, transmission, treatment, symptom, mosquito, and pregnancy.
Research Question 2 - Are tweets on 'Zika virus' relevant?

1.3.1 Goal

We are going to use the user generated contents which are available on twitter to perform disease surveillance. The disease we are interested in is caused by the Zika virus. In this paper, we will analyze how people reacted to Zika virus on twitter and what extent that information can be used for surveillance.

CHAPTER 2

DESIGN DETAILS AND IMPLEMENTATION

Data collection is major and the most tedious part of this project because we required the data from specific geolocation and time duration. For our thesis, we have collected data from twitter in JSON format and later converted into CSV file. We are going to describe how data is fetched, stored, cleaned, processed and classified. Before exploring these processes, let us explain our proposed architecture.

2.1 Proposed Architecture

Our goal is to examine the degree of concern regarding 'Zika virus' by analyzing sentiment polarity of tweets. We are going to pursue following steps to achieve our objective.

Step1 : We are going to extract raw tweets from twitter by using tweepy libraries in python.
Step2 : Then we clean these tweets and remove repeated tweets so that they can be fit for the desired Sentiment analysis algorithm.
Step3 : After preprocessing the data, we are going to use this data set in the algorithm which will classify them as per their polarity.
Step4 : Then we are going to calculate the degree of concern regarding Zika virus.
Since we are going to collect data from twitter so we are going to use twitter application for this purpose. The steps are shown in flowchart given in figure: 2.1.

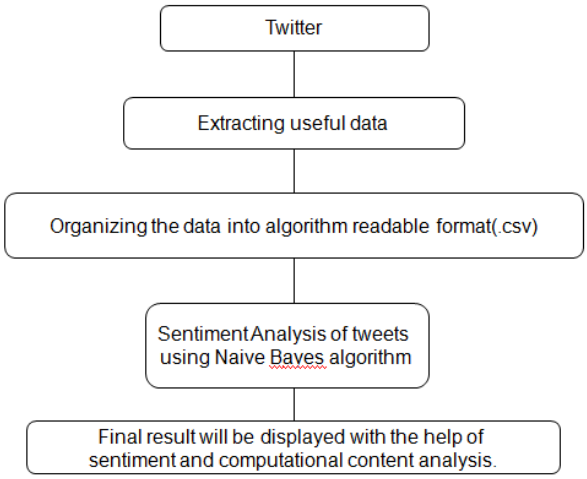

Figure 2.1: Flowchart of proposed activities

2.2 Twitter API

Twitter API is used to extract tweets from twitter. There are two types of twitter APIs: Streaming API and REST API.

2.2.1 Streaming API

Streaming API is used to collect real time tweets.

2.2.2 REST API

Limited data can be fetched using REST API.

2.3 Data Collection

2.3.1 Twitter data

We need a twitter account for using Twitter API.A twitter account can be easily created by filling a sign up form on twitter. Now you will get your login credentials which is used to create our API. We are provided customer secret key, consumer key , access token key and access secret key. These keys are used for authentication purpose while extracting data from twitter.

Since the objective of this thesis is to examine the degree of concern regarding 'Zika virus' by analyzing sentiment polarity of tweets from specific geolocation so we need tweets which contain keywords 'zika', 'zika virus' .We have created python script

for fetching tweets from twitter with keywords 'zika' ,'zika virus' .Before creating this script we required to install Python library known as **tweepy**.

Tweepy is the Python library which helps us to communicate with twitter and use its API to extract data which further can be used in the algorithm to determine sentiment polarity of tweets. Tweepy can be installed by simply using command 'pip install tweepy' in command prompt. In this python script we used customer secret key, consumer key, access token key and access secret key which we are provided with API. First, we create a function that loads the twitter API after authorizing the user.

We have used following functions in our Python script:

1. **load _ api :** This function loads the twitter API after authorizing the user. In this, we use OAth protocol which authorizes the user. OAth provide security and authentication to the user.

2. **tweet _ search :** This function consists of a search string 'query' , max _tweets ,minimum _tweet id, geocode and since id.

3. **get _ tweet _ id :** We use get _tweet _id function in order to get the ID which is considered as 'starting point'.

4. **write _ tweets :** This function writes tweets to a file in JSON format.

5. **main() :**This script continuously searches for tweets. In this part of the script, we can input a specific duration (maximum nine days old) as well as particular phrases(Zika in this thesis) which we want in fetched tweets.

2.4 Data Storage

Once, we start extracting tweets by using Twitter API we required to store that data set as an algorithm readable format. We stored 26,239 tweets from 30 July 2017 to 6 August 2017 having keywords 'Zika' and ' Zika Virus '. Every time a JSON (JavaScript Object Notation) is generated when we ran our python script. These JSON files consist of raw tweets with other information regarding all the particular tweets such as the date when the tweet is done, retweets, geolocation and other information. Since the CSV(Comma Separated Value) format is easily accessible , we converted the JSON data set into CSV format. CSV files can be written/read in less time compared to other formats.

We stored tweets from different dates in different directories on the computer's hard drive. Then we required to preprocess and clean the data before using it in the algorithm for finding out sentiment polarity. So in our next step, we preprocessed the data.

Name	Date modified	Type	Size
zika_2017-07-30.json	07-08-2017 18:20	JSON File	9,834 KB
zika_2017-07-31.json	07-08-2017 18:26	JSON File	13,334 KB
zika_2017-08-01.json	07-08-2017 18:32	JSON File	19,718 KB
zika_2017-08-02.json	07-08-2017 18:39	JSON File	23,491 KB
zika_2017-08-03.json	07-08-2017 18:43	JSON File	14,394 KB
zika_2017-08-04.json	07-08-2017 18:47	JSON File	10,291 KB
zika_2017-08-05.json	07-08-2017 18:50	JSON File	7,855 KB
zika_2017-08-06.json	07-08-2017 18:53	JSON File	8,485 KB

Figure 2.2: .json files containing tweets

2.5 Data Preprocessing

Data obtained from twitter consists of a lot of other information along with tweets which are not fit for extracting features. The raw tweets consist of message along with its metadata .

We required only plain tweet to implement it on our algorithm, so we removed all other metadata such as creation date, language code, location and other information .

Following figures are of one sample raw tweet and its text message. After preprocess-


```
{"created_at": "Wed Aug 02 23:58:41 +0000 2017", "id": 892897472137367553, "id_str": "892897472137367553",
"text": "And so it begins: Texas reports its first case of mosquito-transmitted Zika virus this year.
(https://t.co/FQ1rXsc1Ud)", "truncated": false, "entities": {"hashtags": [], "symbols": [], "user_mentions":
 [], "urls": [{"url": "https://t.co/FQ1rXsc1Ud", "expanded_url": "http://buff.ly/2ho01Qb", "display_url":
"buff.ly/2ho01Qb", "indices": [94, 117]}]}, "metadata": {"iso_language_code": "en", "result_type": "recent"},
"source": "<a href=\"http://bufferapp.com\" rel=\"nofollow\">Buffer</a>", "in_reply_to_status_id": null,
 "in_reply_to_status_id_str": null, "in_reply_to_user_id": null, "in_reply_to_user_id_str": null,
"in_reply_to_screen_name": null, "user": {"id": 2746299102, "id_str": "2746299102", "name": "GPhA",
 "screen_name": "gphabuzz", "location": "Atlanta, GA", "description": "I am the official Twitter of
 the Georgia Pharmacy Association!", "url": "http://t.co/rJeFZ9ash3", "entities": {"url": {"urls":
[{"url": "http://t.co/rJeFZ9ash3", "expanded_url": "http://gphabuzz.com", "display_url": "gphabuzz.com",
"indices": [0, 22]}]}, "description": {"urls": []}}, "protected": false, "followers_count": 620, "friends_count":
 173, "listed_count": 19, "created_at": "Tue Aug 19 19:37:08 +0000 2014", "favourites_count": 53, "utc_offset":
 -25200, "time_zone": "Pacific Time (US & Canada)", "geo_enabled": false, "verified": false, "statuses_count":
 1901, "lang": "en", "contributors_enabled": false, "is_translator": false, "is_translation_enabled": false,
"profile_background_color": "000000", "profile_background_image_url": "http://abs.twimg.com/images/themes/
theme1/bg.png", "profile_background_image_url_https": "https://abs.twimg.com/images/themes/theme1/bg.png",
 "profile_background_tile": false, "profile_image_url": "http://pbs.twimg.com/profile_images/714841210381344768/
anXn-0iS_normal.jpg","profile_image_url_https": "https://pbs.twimg.com/profile_images/714841210381344768/
anXn-0iS_normal.jpg","profile_banner_url": "https://pbs.twimg.com/profile_banners/2746299102/1408997964",
"profile_link_color": "FF9D00", "profile_sidebar_border_color": "000000", "profile_sidebar_fill_color":
"000000", "profile_text_color":"000000", "profile_use_background_image": false, "has_extended_profile":
 false, "default_profile": false,"default_profile_image": false, "following": false, "follow_request_sent":
 false, "notifications": false,"translator_type": "none"}, "geo": null, "coordinates": null, "place": null,
 "contributors": null,"is_quote_status": false, "retweet_count": 0, "favorite_count": 0, "favorited": false,
 "retweeted":false, "possibly_sensitive": false, "lang": "en"}
```


Figure 2.3: Sample of one raw tweet


```
"text": "And so it begins: Texas reports its first case of mosquito-transmitted Zika virus this year.
```


Figure 2.4: Text message of the tweet

ing the data, it is ready for our next step which is to use this pre-processed data set on the algorithm to classify them into different polarity groups (positive, negative, neutral).

2.6 Classification

Sentiment analysis or opinion mining is the process through which we decide any write up into three polarity classes which are positive, negative and neutral.For classifying the tweets into different polarity classes there are many techniques and algorithm available. So we classified the tweets in different classes (positive, negative, neutral) by using one of the techniques 'Naive - Bayes Classifier'.

CHAPTER 3

RESULTS AND DISCUSSION

In this chapter, we are going to present various result that we have got from our implementation.

3.1 Tweets Collected

We collected tweets by using Twitter API. A .json file is generated each time when we ran our Python script for tweets fetching. This file consists of original text message of tweets as well as much other information. We removed all other unnecessary information and created a data set which consists only original text. A sample file of tweets is shown in Figure:3.1 We collected 26,239 tweets between 30th July 2017 and

```
What do you do right now? \u2014 Sleep w eat https://t.co/4X9Qn07Bz2
RT @WakingTimes: Zika: Brazil Admits It\u2019s Not the Virus- Plz ReTweet #wakingtimes https://t.co/KQmFqNtnVH
RT @WakingTimes: Zika: Brazil Admits It\u2019s Not the Virus- Plz ReTweet #wakingtimes https://t.co/KQmFqNtnVH
#wheresjojo @WorstOfAnything Zika the best birth control! #sorrynotsorry #nofilter #standup\u2026 https://t.co/8eUeiV2B1w
JOHN GOODMAN SWEARS HE DOESN'T HAVE THE ZIKA VIRUS! https://t.co/WzqH1HbWv6
RT @_Infidoll_: 1947 Rockefeller Patent Shows Origins Of Zika &amp; What About Those Genetically Modified Mosquitoes https://t.co/VVdh6NSiel vi\u2026
RT @_Infidoll_: 1947 Rockefeller Patent Shows Origins Of Zika &amp; What About Those Genetically Modified Mosquitoes https://t.co/VVdh6NSiel vi\u2026
I'm gonna make an app that lets you track how many mosquitos are around you and which ones have Zika/West Nile Virus
Not that I trust the Rockerfellers but zika was first found in wild monkeys. \ud83d\ude15 https://t.co/diBifoBoFf
Not that I trust the Rockerfellers but zika was first found in wild monkeys. \ud83d\ude15 https://t.co/diBifoBoFf
Researchers resurrect old antibiotic molecule in hopes of treating Zika https://t.co/oWI5FYcN2d
Scientists Are Closer To A Zika Vaccine That Protects Babies Of \u2026 : https://t.co/F4GvpSUUg7
RT @Zika_LSHTM: Case of #Zika virus
RT @Zika_LSHTM: Case of #Zika virus
RT @WakingTimes: Zika: Brazil Admits It\u2019s Not the Virus- Plz ReTweet #wakingtimes https://t.co/KQmFqNtnVH
RT @WakingTimes: Zika: Brazil Admits It\u2019s Not the Virus- Plz ReTweet #wakingtimes https://t.co/KQmFqNtnVH
Ashton the Great died from Zika Virus in 1802.
We should be careful about female mosquitos for preventing #zika and other #virus!\nhttps://t.co/CLsDvkROtA
We should be careful about female mosquitos for preventing #zika and other #virus!\nhttps://t.co/CLsDvkROtA
Come out with a years worth of pink eye and zika \ud83e\udd26\ud83c\udffd\u200d\u2642\ufe0f https://t.co/kpK19Kb1JQ
Come out with a years worth of pink eye and zika \ud83e\udd26\ud83c\udffd\u200d\u2642\ufe0f https://t.co/kpK19Kb1JQ
RT @WakingTimes: Zika: Brazil Admits It\u2019s Not the Virus- Plz ReTweet #wakingtimes https://t.co/KQmFqNtnVH
RT @WakingTimes: Zika: Brazil Admits It\u2019s Not the Virus- Plz ReTweet #wakingtimes https://t.co/KQmFqNtnVH
Zika: Brazil Admits It\u2019s Not the Virus- Plz ReTweet #wakingtimes https://t.co/KQmFqNtnVH
It's that time of year! Stock up on Skin So Soft products today at https://t.co/Q49413J917 #travel #zika #camping\u2026 https://t.co/ckReXM0bQE
RT @greg_folkers: .@NatureComms : A peptide-based viral inactivator inhibits #Zika virus infection in pregnant mice and fetuses https://t.c\u2026
RT @greg_folkers: .@NatureComms : A peptide-based viral inactivator inhibits #Zika virus infection in pregnant mice and fetuses https://t.c\u2026
@FoxNews @JudgeJeanine At least Dems &amp; indeps have minds to loose. U on the other hand must have been born with zik\u2026 https://t.co/594bmF4Nfr
@FearTheKings2 @Smackin_HosT you kids give me ZIKA EBOLA AIDS STAGE 10 CANCER BRAIN DAMAGE BLOOD LOSS ETC
Here's another reason to provide healthcare to all Americans. Also
Here's another reason to provide healthcare to all Americans. Also
I liked a @YouTube video https://t.co/YmoY34nDem Hope Solo Gets Taunted By Thousands Screaming 'Zika!' at Olympic Match In Rio
\Lil Uzi Sounds Like Someone Is Losing Blood At Uncontrollable Rate And No One Can Help You\ -@jess_zika \ud83d\ude02
\The human Zika virus that is known as Kellyanne Conway...\ I want all news outlets to use this. It's on me.
What is zika virus and how is it vanished? by Imtiyaz Ali https://t.co/ducmNmdZRQ
@FoxNews Hey FL - Forget about Zika -Call in the gators #BearSprayForAll \ud83c\udf40
```

Figure 3.1: Sample collected tweets

6th August 2017. The number of tweets collected per day is shown in Figure :3.1

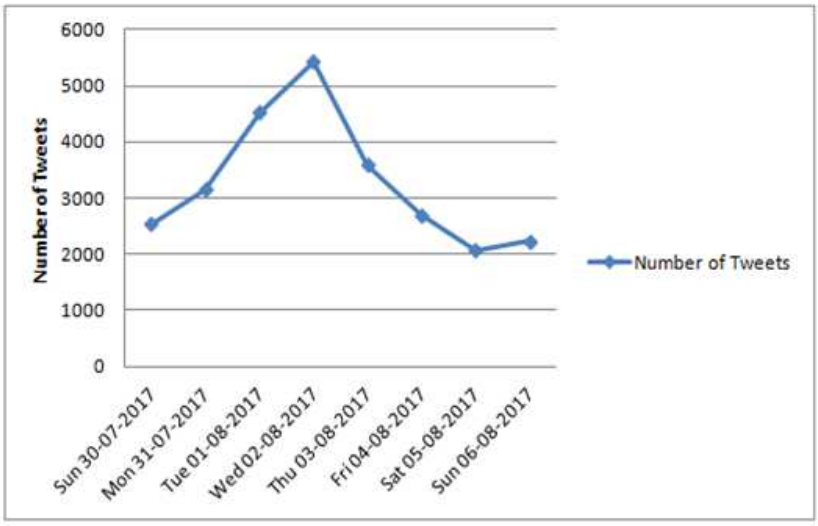

Figure 3.2: Collected tweets per day

3.2 Twitter Data Analysis

For our thesis, we collected tweets which consist keywords 'Zika' continuously for eight days. We pre-process our dataset so that we can use them for sentiment classification. We are going to pass our dataset to the Naive Bayes classifier. Sample tweets with its polarity are shown in Figure:3.2 When we ran our Python script for sentiment

3111	"text": ".@CDCgov updates guidance for providers caring for #pregnant #women w/ possible #Zika exposure\u	0.166667	0.75	Positive
3112	"text": "No bail\u00e3o	0	0	Neutral
3113	"text": "RT @Ruffles_Oficial: Maratona de leitura pra relembrar os 20 anos do bruxo mais zika de todos os tem	0	0	Neutral
3114	"text": "@kfrydl hi	0	0	Neutral
3115	"text": "RT @Ruffles_Oficial: Maratona de leitura pra relembrar os 20 anos do bruxo mais zika de todos os tem	0	0	Neutral
3116	"text": "@Yes_Zika Okayyyy I'll text you around 9 and see what is up!"	0	0	Neutral
3117	"text": "@zonumonurb eita zika"	0	0	Neutral
3118	"text": "@BadKidOscar I wont be able to at 8. Just come at like 9 or 10"	0.5	0.625	Positive
3119	"text": "RT @Ruffles_Oficial: Maratona de leitura pra relembrar os 20 anos do bruxo mais zika de todos os tem	0	0	Neutral
3120	"text": "RT @Ruffles_Oficial: Maratona de leitura pra relembrar os 20 anos do bruxo mais zika de todos os tem	0	0	Neutral
3121	"text": "RT @thsefudeu: Look pra visitar a amiga com zika https://t.co/V3K20Ys43u"	0	0	Neutral
3122	"text": "RT @valeriejanz: Video: What you need to know about the Zika virus https://t.co/dQAXjpN4KA via ma	0	0	Neutral
3123	"text": "eu preto Zika	0	0	Neutral
3124	"text": "RT @APPCPenn: What happens to public trust in #science after news of scientific breakthrough like #Z	0	0.066667	Neutral
3125	"text": "C\u00f3mo el cambio clim\u00e1tico ayudar\u00e1 a predecir virus como el Zika y el \u00c9bola #saluc	0	0	Neutral
3126	"text": "10% Off This Week - Awesome Bug Repellent #repellent #bmrtg #zika #discount #noseeum https://t.c	-0.26667	1	Negative
3127	"text": "10% Off This Week - Awesome Bug Repellent #repellent #bmrtg #zika #discount #noseeum https://t.c	-0.26667	1	Negative
3128	"text": "10% Off This Week - Awesome Bug Repellent #repellent #bmrtg #zika #discount #noseeum https://t.c	-0.26667	1	Negative
3129	"text": "RT @Ruffles_Oficial: Maratona de leitura pra relembrar os 20 anos do bruxo mais zika de todos os tem	0	0	Neutral
3130	"text": "@WladimirJara En Chile no hay dengue ni zika genio. Eso es principalmente por el aedes egypty	0	0	Neutral

Figure 3.3: Filtered tweets with output

polarity on the dataset which consists of 26239, it classified data into their polarity and printed out result in terms of numbers as well as percentage. In IPython console it printed the result which is shown in Figure:3.2 The Python script classified 20302 tweets as neutral, 4603 as positive and 1334 as negative. In a similar study done in 2016 [9] they found that as much as 59 % of the tweets on Zika virus was negative,

```
In [13]: runfile('H:/STUDY/BTP/CODE/123.py', wdir='H:/STUDY/BTP/CODE')
4603 1334 20302
Positive tweets percentage: 17.542589275505925 %
Negative tweets percentage: 5.08403521475666 %
Netural tweets percentage: 77.37337550973741 %
```

Figure 3.4: In Python console printed the result

Polarity	Number of tweets	Percentage of tweets(%)
Positive	4603	17.54
Negative	1334	5.08
Neutral	20302	77.37

Table 3.1: Table of Statistical data

and that was the time when WHO (World Health Organization) declared emergency [11, 18] because there were estimated millions of cases from around the world was reported during 2015 - 2016. Therefore there was very high level of concern among society regarding this epidemic and hence it was visible in the report where 59 % of tweets were negative. In our study, we found that only 5.08 % tweets are negative because we have done our study at such a time when Zika virus is no longer a matter of high concern among people.

So by comparing these two results, it is clear that Twitter could play a significant role in case of any epidemic or public health emergency to decide the level of concern among people by using sentiment analysis. We have analyzed all the tweets by dividing them into six categories which are Symptoms, Treatment, Prevention, Transmission, Mosquito, and Pregnancy. We included two other classes along with four diseases characteristics. Since Zika virus is spread through mosquitoes and a pregnant mother could pass the Zika virus to her infant. These are the two main reasons we included two other classes Mosquito and Pregnancy. We found that very high number of people were tweeting about the four diseases characteristics, mosquito, and pregnancy which shows that very large number of people were tweeting relevant to this epidemic.

After analyzing the data we found that most numbers of people were twitting about transmission, mosquito, and prevention. Not many people were twitting about treatment, the reason for that is there is not any convincing treatment for Zika virus. Sample tweets from all these categories are shown in table :3.2.

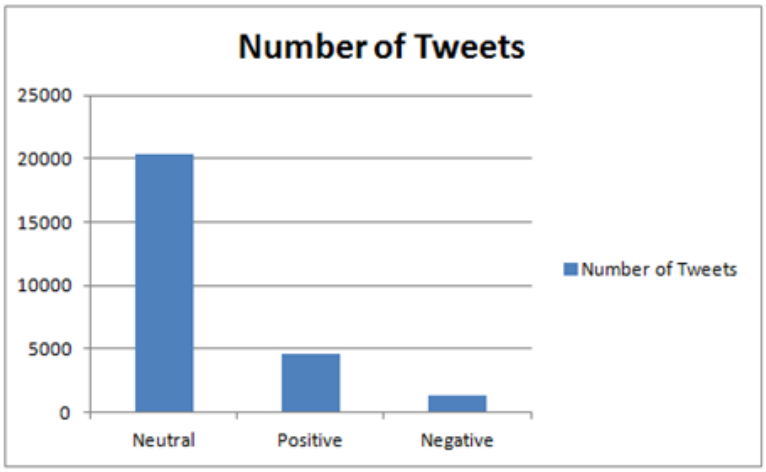

Figure 3.5: Graph between polarity and number of tweets

Category	Tweet
TREATMENT	KiltronX has Pesticide Free Treatment that kills them! # Mosquito Season Is Spreading # Zika and West Nile Across U.S.
PREVENTION	What other insect-borne diseases are occurring in the Zika areas? What is main method of prevention
SYMPTOMS	Learn where symptoms indicative of Zika are in the world with # Kidenga . # zika # dengue # healthapp
TRANSMISSION	Zika May Have a Startlingly High Sexual Transmission Rate
MOSQUITO	We should be careful about female mosquitos for preventing #zika and other # virus
PREGNANCY	Health agency clarifies Zika pregnancy guidance... Again

Table 3.2: Table of Tweets

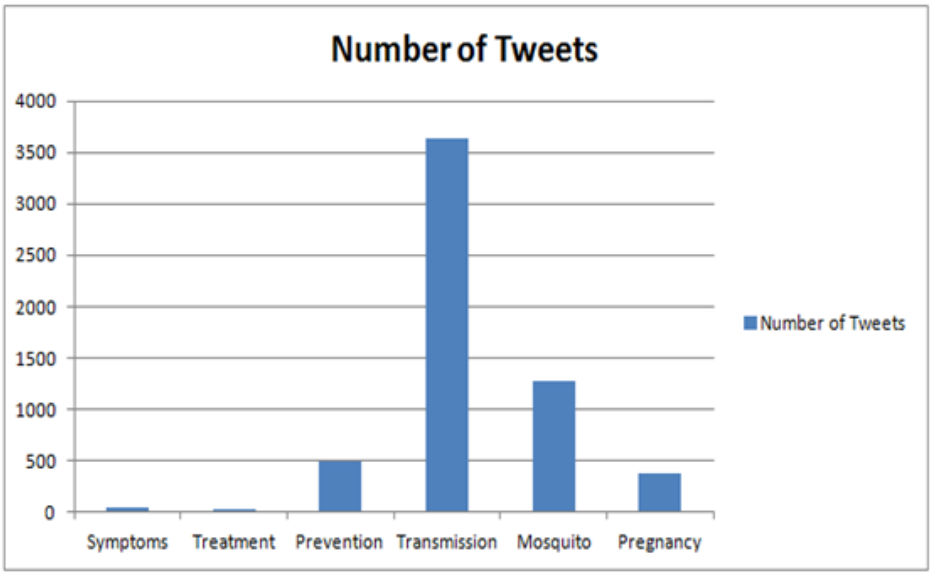

Figure 3.6: Number of tweets in each categorization after running all tweets

CHAPTER 4

CONCLUSION AND FUTURE SCOPE

4.1 Conclusion

Sentiment analysis on social media data is used to is used to find out people's opinion on particular topic in terms of polarity of contents which people share on social media like twitter .

Twitter is a large source of information about people's view which makes it one of the best sources for doing sentiment analysis. We performed sentiment analysis on 26239 tweets from the United States of America (USA) and some parts of Canada . on Zika virus.All these tweets were done between 31st July 2017 and 6th August 2017.We found that only 5.08% of tweets were negative which shows that during this period the concern regarding Zika virus is very less among people as expected. Since even World Health Organization (WHO) announced earlier that Zika Virus was no longer a matter of serious health crisis the result which we got was well desired. But according to a similar study done in 2016 shows that as many as 59% tweets were negative which shows very high degree of concern and the reason behind such result was millions of cases regarding Zika virus was reported around the globe and also World Health Organization (WHO) during that period announced Zika virus to be next big health crisis . So these very much accurate and result signifies that Twitter data can be a very good source to find out public concern at the time of any health crisis.

Six phrase classes which we chose for analysis showed that many people are well aware of the Zika virus and many people twitted about mosquito and pregnancy which are the primary source of spread of Zika virus. It is not necessary that our method and algorithm can be used only in case of Zika virus. It can be utilized for various other purposes depending upon the tweets which we collect using specific keywords.

4.2 Future Scope

Future scope which could be added to our research are:
1. We can initiate to work in multiple languages to provide analysis to more locations where multiple languages are spoken.
2. We can improve our system to filter out sentences which are not relevant to the topic but contains the common keywords of tweet collection.

REFERENCES

[1] Bhattacharya, S., Tran, H. and Srinivasan, P.: 2012, Discovering health beliefs in twitter., *AAAI Fall Symposium: Information Retrieval and Knowledge Discovery in Biomedical Text*.

[2] Bird, S.: 2006a, Nltk: the natural language toolkit, *Proceedings of the COLING/ACL on Interactive presentation sessions*, Association for Computational Linguistics, pp. 69–72.

[3] Bird, S.: 2006b, Nltk: the natural language toolkit, *Proceedings of the COLING/ACL on Interactive presentation sessions*, Association for Computational Linguistics, pp. 69–72.

[4] Cao-Lormeau, V.-M., Blake, A., Mons, S., Lastère, S., Roche, C., Vanhomwegen, J., Dub, T., Baudouin, L., Teissier, A., Larre, P. et al.: 2016, Guillain-barré syndrome outbreak associated with zika virus infection in french polynesia: a case-control study, *The Lancet* **387**(10027), 1531–1539.

[5] Go, A., Huang, L. and Bhayani, R.: 2009, Twitter sentiment analysis, *Entropy* **17**, 252.

[6] Madhukar, G. V., Bhausaheb, D. K., Babasaheb, K. K., Tukaram, D. R. and Balkrishna, S. S.: 2016, Zika virus infection: An overview.

[7] Malone, R. W., Homan, J., Callahan, M. V., Glasspool-Malone, J., Damodaran, L., Schneider, A. D. B., Zimler, R., Talton, J., Cobb, R. R., Ruzic, I. et al.: 2016, Zika virus: medical countermeasure development challenges, *PLoS neglected tropical diseases* **10**(3), e0004530.

[8] McKinney, W.: 2012, *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*, " O'Reilly Media, Inc.".

[9] Miller, M., Banerjee, D., Muppalla, R., Romine, D., Sheth, D. et al.: 2017, What are people tweeting about zika? an exploratory study concerning symptoms, treatment, transmission, and prevention, *arXiv preprint arXiv:1701.07490* .

[10] Murphy, K. P.: 2006, Naive bayes classifiers, *University of British Columbia* .

[11] Organization, W. H. et al.: 2016, Who director-general summarizes the outcome of the emergency committee regarding clusters of microcephaly and guillain-barré syndrome, *Saudi medical journal* **37**(3), 334.

[12] Pang, B., Lee, L. and Vaithyanathan, S.: 2002, Thumbs up?: sentiment classification using machine learning techniques, *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, Association for Computational Linguistics, pp. 79–86.

[13] Pang, B., Lee, L. et al.: 2008, Opinion mining and sentiment analysis, *Foundations and Trends® in Information Retrieval* **2**(1–2), 1–135.

[14] Paul, M. J. and Dredze, M.: 2011, You are what you tweet: Analyzing twitter for public health., *Icwsm* **20**, 265–272.

[15] Petersen, L. R., Jamieson, D. J., Powers, A. M. and Honein, M. A.: 2016, Zika virus, *New England Journal of Medicine* **374**(16), 1552–1563.

[16] Purohit, H., Banerjee, T., Hampton, A., Shalin, V. L., Bhandutia, N. and Sheth, A. P.: 2015, Gender-based violence in 140 characters or fewer: A# bigdata case study of twitter, *arXiv preprint arXiv:1503.02086* .

[17] Sikka, V., Chattu, V. K., Popli, R. K., Galwankar, S. C., Kelkar, D., Sawicki, S. G., Stawicki, S. P. and Papadimos, T. J.: 2016a, The emergence of zika virus as a global health security threat: a review and a consensus statement of the indusem joint working group (jwg), *Journal of global infectious diseases* **8**(1), 3.

[18] Sikka, V., Chattu, V. K., Popli, R. K., Galwankar, S. C., Kelkar, D., Sawicki, S. G., Stawicki, S. P. and Papadimos, T. J.: 2016b, The emergence of zika virus as a global health security threat: a review and a consensus statement of the indusem joint working group (jwg), *Journal of global infectious diseases* **8**(1), 3.

[19] Wang, H., Can, D., Kazemzadeh, A., Bar, F. and Narayanan, S.: 2012, A system for real-time twitter sentiment analysis of 2012 us presidential election cycle, *Proceedings of the ACL 2012 System Demonstrations*, Association for Computational Linguistics, pp. 115–120.

[20] Zhang, H.: 2004, The optimality of naive bayes, *AA* **1**(2), 3.