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:

**Proposed Architecture:**

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

**FIGURE 1; FLOWCHART**

**Step1:** We are going to extract raw tweets from twitter by using tweepy libraries in python.

**Step 2:T**hen we clean these tweets and remove repeated tweets so that they can be fit for the desired Sentiment analysis algorithm.

**Step 3:** After preprocessing the data we are going to use this data set in the algorithm which will classify them as per their polarity .

**Step 4:**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 purose.

**2. Twitter API:**

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

**Streaming API-** Streaming API is used to collect real time tweets .

**REST API -** Limited data can be fetched using REST API .

**3. DATA COLLECTION**

**3.1 Twitter data:** We need atwitter 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 degree of concern regarding 'Zika virus' by analyzing sentiment polarity of tweets from specific geolocation so we need tweets which contains 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 function that loads the twitter API after authorizing the user .

We have used following functions in our Python script:

**load\_api :** Thisfunction loads the twitter API after authorizing the **user.** In this we use OAth protocol which authorizes the user. OAth provide security and authontication to user.

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

**get\_tweet\_id :**  we use get\_tweet\_id function in order to get the ID which is considered as 'staring point'

**write\_tweets :**This function writes tweets to a file in JSON format .

**main() :** This script continuously searches for tweets . In this part of the script we can input the specific duration (maximum 9 days old) as well as specific phrase(Zika in this thesis ) which we want in fetched tweets.

**SCREEN SHOT OF CODE AND DATASET IN JSON**

**4.DATA STORAGE**

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

**figure- screen shot of the tweet files from different date**

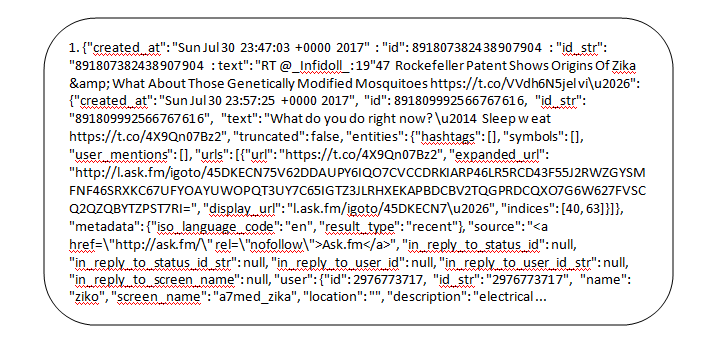
We stored tweets from different dates in different directories in the computer hard drive . Then we required to pre-process and clean the data before using it on algorithm for finding out sentiment polarity . So in our next step we pre-processed the data .

**DATA PRE-PROCESSING**

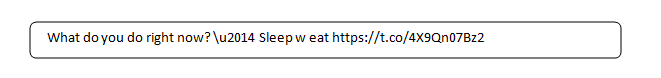
Data obtained from twitter consists of a lot of other informationalong with tweets which is not fit for extracting features. The raw tweets consist of message along with its metadata .

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

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



one raw tweet.



text message of the tweet.

We removed all the stop words ( like is , this ,the , he , a etc. ) from the main message of the tweet by using various NLTK functions .

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

**6. CLASSIFICATION**

Sentiment analysis or opinion mining is the process through which we decide any write up in to 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' .

