Journal of Computational Science 25 (2018) 406–415

Contents lists available at ScienceDirect

Journal of Computational Science

journal homepage: www.elsevier.com/locate/jocs

Effective surveillance and predictive mapping of mosquito-borne diseases using social media

Vinay Kumar Jain*, Shishir Kumar

Department of Computer Science & Engineering, Jaypee University of Engineering & Guna (M.P.), India

ARTICLE INFO

Article history:
Received 30 December 2016
Received in revised form 29 June 2017
Accepted 6 July 2017
Available online 8 July 2017

Keywords:
Mosquito-borne diseases
Twitter
Medical informatics
Classification
LDA
Kernel density
Sentiment analysis

ABSTRACT

Healthcare Emergency Management involves preventing, handling, organizing and controlling of specific events and in response to emergency situations. A social media based mosquito-borne disease surveillance and outbreak management using spatial and temporal information which help in identification, characterization, and modeling of user behavioral patterns on the web have been presented through this paper. The proposed predictive mapping based on geo-tagging data has a significant impact on preventing and tracking mosquito-borne disease in the specific area with limited resources. The tracking of real-time public sentiments provides an early discovery or alarming related to outbreak. Latent Dirichlet allocation (LDA) based topic modeling techniques have been applied to filter out relevant topics related to symptoms, prevention and fear. The two steps fine-grained classifications of data have been performed using Naive Bayes and Support Vector machine. The proposed framework focused on alternative methods of analysis and visualization of user's opinions that do not depend upon the assumption of normality. A novel intelligent surveillance process model has been presented which help government agencies for proper management of time and resources. The utilization of standard kernel density estimate (KDE) with important factors derived from Twitter and RSS feeds have been presented for predictive mapping. The model uses latent Dirichlet allocation for identification of coherent topics from collected data set at a particular interval. This model has been applied to predict the occurrence of mosquito-borne disease in India.

1. Introduction

The epidemics related to Mosquito-Borne Diseases (MBD) are unrelenting economic and health burden around the globe. These epidemics now become a global threat in endemic and non-endemic regions. The effects of Mosquito-Borne Diseases (MBD) have will continue for several decades resulted in thousands of deaths every year. The early detection and monitoring system for outbreaks has been established but due to inappropriate information, this system does not work efficiently [1]. The disease outbreaks often develop rapidly, are difficult or impossible to predict and cause a disproportionately high burden due to the lack of response capabilities. As a result of the clear importance of disease outbreaks to wider control efforts, research agendas and subsequent policy guidelines have heavily focused on methods to predict outbreaks.The outbreak softens develop rapidly, are difficult or impossible to predict and cause a disproportionately high burden due to the lack of response capabilities [2]. Hence, enhanced reconnaissance through quick reaction measures against Mosquito-Borne Diseases is a long-standing precept to different wellbeing frameworks around the globe. They are desperately required to moderating the officially overwhelming weight on those wellbeing frameworks and restricting further spread of mosquito-borne diseases inside topographical areas [3]. The flare-ups can apply expansive weights on general well-being frameworks, as healing centers and outpatient facilities get to be overpowered by the surge in genuine dengue positive cases, and additionally other febrile diseases [4].These weights are intensified by asset restricted or powerless observation frameworks. The capacity to anticipate flare-ups with a liberal slack time ought to empower general wellbeing frameworks to react all the more effectively through the auspicious allotment of assets [5,6]. Social media is transforming itself into a powerful and adaptable tool that many public and private organizations increasingly seek to understand. The social media is a global digital community where individuals stay connected with each other by direct and split-second electronic exchange of information constantly throughout the day [7,8]. While forms of social media have been around for ages, this is rev-

* Corresponding author.
E-mail addresses: vinay2588@gmail.com (V.K. Jain), dr.shishir@juet.ac.in (S. Kumar).

https://doi.org/10.1016/j.jocs.2017.07.003
1877-7503/© 2017 Elsevier B.V. All rights reserved.

olutionary, since people now communicate instantaneously across national borders and oceans in mere seconds [9].

Traditional methods of detection rely on hierarchical, bureaucratic, health-care system structures that add to time delays in detection that cannot usually keep up with the speed at which an infectious disease is spread [8]. However, the process of collecting, integrating, and analyzing disease-related information from diverse web-based channels becomes more difficult and challenging as social media platforms proliferate and the amount of data quickly multiply [9]. Further, information from multiple sources brings up new challenges for information reliability and validity.

The Geographic Information Systems (GIS) provides geo-locations which are widely used in real time application for identification of affected locations. In the proposed work geo-locations are extracted from Twitter. These geo-locations help to determine the future projection of rate of spreading of diseases in nearby locations. With this information, for example, government health agencies can take the rapid decision in most heavily affected locations. They can easily identify specific neighborhoods and communities that are of highest importance for urgent interventional care. Social media is one of the main resources which directly linked with the community suffered from a particular disease.

The proposed system can be utilized for tracking and preventing mosquito-borne disease in developing countries with limited resources. The framework provides a social media based disease surveillance and outbreak management which help in identification, characterization, and modeling of user interests and behavioral patterns on the web.

The proposed surveillance process model has been used for detecting diffusion of mosquito-borne diseases in India that has been occurred in 2016. This framework utilized integration of social media analytics and Geographic Information System (GIS) method for predicting the outbreak. Three major research objectives have been presented through this paper: (a) To investigate the spatiotemporal relationship between the advancement of mosquito-borne diseases in India using Twitter posts and RSS feeds; (b) Examine the public sentiments related to mosquito-borne diseases; and (c) To show the spread of the mosquito-borne diseases in space and time.

2. Related work

In the previous 10 years, research articles interfacing illness diseases connected with web-based social networking have expanded in number because of the expansion in accessibility of real-time data from different geo-locations [10]. Healthcare data extraction can fabricate a database with the data on a given diseases drawn from online resources, for example, on the web medicinal news, biomedical writing, blogs, forums and social media platforms. Numerous websites act as potential channels for individuals to discuss symptoms and share their geographical location, so governments and private enterprises have explored various models in detecting and tracking potential pandemics [11,12]. The literature identified major gaps and biases in utilizing social media in the area of public health practices.

Authors developed social media based early warning systems for mosquito-borne diseases such as Lowe et al. [13] and Racloz et al. [14] developed efficient early warning and surveillance system for dengue fever. Hay et al. [15] and Thomson et al. [16] identified the important aspect of malaria with surveillance. Natural Language Processing (NLP) techniques such as corpus-based and sentiment analysis have been providing a rich source of information for detecting and forecasting disease outbreak in all around the world [17]. Chew and Eisenach [8] used specific keywords related to outbreak detection in 2009 H1N1 pandemic.Hu et al. [9] used web services provided by Google related to influenza epidemic using specific keywords. Lampos and Cristianin [10] used content based methods with statistical methods to monitor and measure public perceptions. They also analyzed levels H1N1 pandemic. Chunara et al. [11] detected cholera outbreak using Twitter. Aramaki et al. used Support Vector Machine for predicting influenza rates in Japan [12]. Stewart et al. [18] developed a real-time data analysis of disease using social media with an early warning system. Bodnar et al. [19] applied various classification techniques for detecting influenza. Parket et al. developed a framework to tracks the levels using trends extracted from Twitter [20]. Jain and Kumar [21] has tracked levels of Influenza-A (H1N1) during 2015.

The Internet has changed effective real-time healthcare monitoring system for epidemic diseases insight [22]. The expanded recurrence of Internet use for gaining healthcare data has added to the ascent of online early detection systems for infectious diseases [23]. The primary idea is that ailment related data is recovered from an extensive variety of accessible real-time electronic information sources, which play basic parts in the distinguishing proof of early healthcare events and situational readiness [24].

3. Mosquito-borne diseases

Every year, there are roughly a large number of instances of mosquito-borne infections has been accounted for around the world [25]. Dengue is the most well-known mosquito-borne infection with around 60–100 million cases every year [26]. This mosquito-borne disease spreads in developing countries due to inadequate water management, substandard housing, airborne travel, immigration and deteriorating disease prevention programs.

3.1. Dengue

The dengue infection is transmitted from human to human by the bites of contaminated female mosquitoes, for example, Aedes aegypti and Aedes albopictus. These species are also responsible for transmitting other mosquito-borne viruses; including dengue and dengue hemorrhagic fever [25–27]. Infected human feels illness usually between 4 and 8 days but in many reported cases it can range from 2 to 12 days. The medical science specialists and researchers are creating many methodologies that investigate the freshly discovered capacities of distinguishing mosquito-borne diseases. The worldwide occurrence of dengue infection (DENV) contamination has expanded drastically, with pestilences including more serious cases so satisfactory observation to identify early cases is urgent. The utilization of traditional method for surveillance of insect's population is used to predict mosquito-borne diseases cases [27].

3.2. Chikungunya

Chikungunya disease is a viral infectious disease transmitted by the bite of infected mosquitoes same as concerning dengue [25–27]. Chikungunya infection is normally found in tropics and consequently, Chikungunya is overwhelmingly found in Asian nations. Chikungunya side effects incorporate serious and persevering joint agony, body rash, cerebral pain, and fever. The starting side effects are like dengue fever and similar symptoms [26]. Surveillance of insect's population is used of dengue to predict a rough estimate of cases.

3.3. Malaria

Malaria is a disease caused by parasites that are transmitted to human through the bites Anopheles mosquitoes [25–27]. Malaria

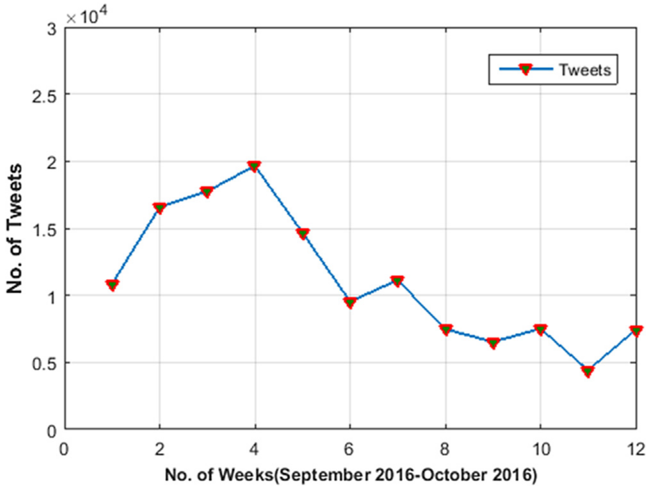

Fig. 1. No. of tweets collected weekly (September 2016–November 2016).

is an intense febrile sickness. In a non-immune individual, symptoms appear 7 days or more after the infective mosquito bite. The common symptoms are chill, fever, headache, and vomiting [25]. Developing countries have a high burden of malaria due to weak surveillance systems and illiteracy. The tradition system lacks in getting real-time disease distribution and trends due to this it is difficult in early decision-making [27].

4. Data acquisition and processing

In this section, the technique for data collection and pre-processing has been presented. The data has been collected from two sources i.e. Twitter and news articles. The important variables for data collection from Twitter are keywords, which helps in identification of relevant tweets. For an effective surveillance system, it is important to fetch relevant tweets which indicate the presence of mosquito-borne diseases or its symptoms. Twitter API [28] has been used for fetching data using relevant keywords and medical science terms. The keywords collection methodology is based Jain and Kumar [17] which gives dynamic words which are well known amid during a specific day and related to general trends and public feelings. Some of the trending keywords are:

Keywords:{#Dengue,# Chikungunya, #Zika,#DengueFever,#yellow fever#Dengue virus,#Zika virus,#Flu,#Swine Flu,#Fluvirus}

Fig. 1 illustrated the no. of tweets collected in between September 2016 to November 2016. Every word contains in a tweet is important in decision making, so pre-processing of these tweets is an important task because these messages are full of misspellings, slang, and words from other languages. In order to tackle the problems with the noise in texts important text processing techniques have been applied such as tokenization, stemming, lemmatization, stop words removal, dimensionality reduction, feature weighting, frequency based methods etc [29].

5. Mining twitter streams

Data analyses on tweet attributes are important in order to understand the hidden patterns and for preparing effective datasets which are used for experimental studies. Hence, the first experiments will lead towards the classification of tweets in three major classes and the second experiment will lead towards sentiment analysis.

Table 1
Results of ten-fold cross-validation for disease related tweets/irrelevant tweets classification.

Category	Naïve Bayes	SVM
Disease related tweets	65.6	68.3
Irrelevant Tweets	64.8	69.2

5.1. *Tweets classification*

The essential thought is that tweets are spoken to as arbitrary blends of words related to symptoms, fear, prevention, and care. In this task three main classes symptoms, fear, and prevention have been created using bag-of-words taking from multiple sources [25–27]. The suitable class has been assigned to a tweet based on suitable lexicons. In order to reach the goals stated, the system activities can be characterized by the following sequence of activities:

Firstly, the dataset is divided into two basic classes, namely, diseases related tweets and irrelevant tweets using Support Vector Machine (SVM) and Naive Bayes(NB). Secondly, effective fine-grained classification of relevant tweets into three categories main classes symptoms, fear, and prevention has been performed using SVM and NB and is presented in Fig. 2.

For evaluation purposes, a baseline system has been developed. The system that counts the feature words of every category in a tweet. The category with the largest number of feature words existing in a tweet has been assigned to it. For obtaining prior knowledge about feature words, seeds words related to mosquito-borne disease have been extracted from multiple resources such as CDC, AMCA,NVBDCP [25–27]. Table 1 lists the important feature words considered for classification in two three categories.

A comparative performance evaluation of Naïve Bayes(NB) and Support Vector Machine (SVM) in terms of correctly classify classes containing tweets has been examined. The results are explained in terms of precision, recall, accuracy, and F-measure and represented in Table 1. For building the training and testing sets, 10,000 random tweets have been extracted for training and 1000 tweets are extracted for testing. Ten-fold cross-validation experiments have been conducted using the NB and SVM. Precision and recall have been calculated using Multi-class classification of text techniques which are presented in the following references [11,12,21]. The F-measure provides the overall performance of a classifier and is calculated using the following formula given by Eq. (1).

$$\mathrm{F} - \mathrm{measure} = \frac{2\,(\mathrm{Precision})\,(\mathrm{Recall})}{\mathrm{Precision} + \mathrm{Recall}} \tag{1}$$

On Comparing, the performance of the classifier, SVM classifier was found better than the Naïve Bayes classifier. The highest accuracy achieved was 69.20%.

NB and SVM method has been used for predicting the class of the tweets in the data set. The classifiers with features of each class have been trained using seed words taken from multiple sources [25–27]. In all experiments, a tweet was represented by a vector having values which indicate the occurrence of each feature. Table 2 present results of ten-fold cross-validation experiments conducted using the Naïve Bayes and SVM implementation.

The performance using the SVM classifier was found better than Naïve Bayes classifier. The text classification deals with high dimensionality of feature space because of it various learning algorithms do not work with high dimensional feature space.

5.2. *Sentiment analysis*

Sentiment analyses are related to monitors public perceptions about events. It helps in analyzing what people think about events

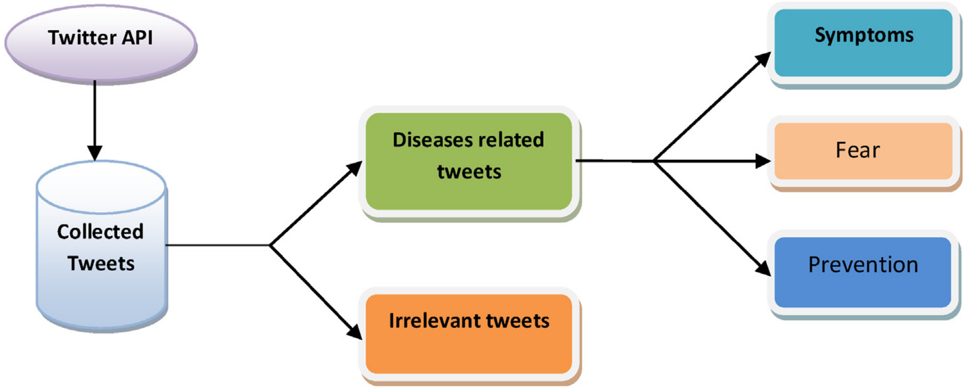

Fig. 2. Proposed tweet classification method.

Table 2
Results of fine-grained classification using Naive Bayes and SVM.

Class	Naive Bayes			SVM		
	Precision	*Recall*	*F-measure*	*Precision*	*Recall*	*F-measure*
Symptoms	0.713	0.367	0.485	0.649	0.297	0.408
Prevention	0.426	0.349	0.384	0.401	0.372	0.386
Fear	0.314	0.312	0.313	0.314	0.332	0.323

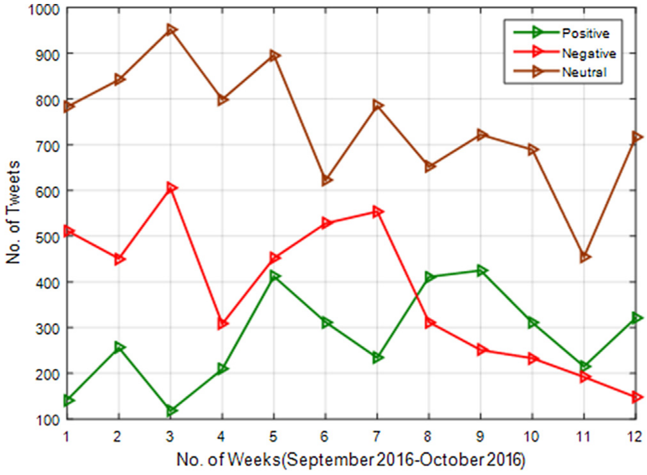

Fig. 3. Sentiment classification by tweets polarity.

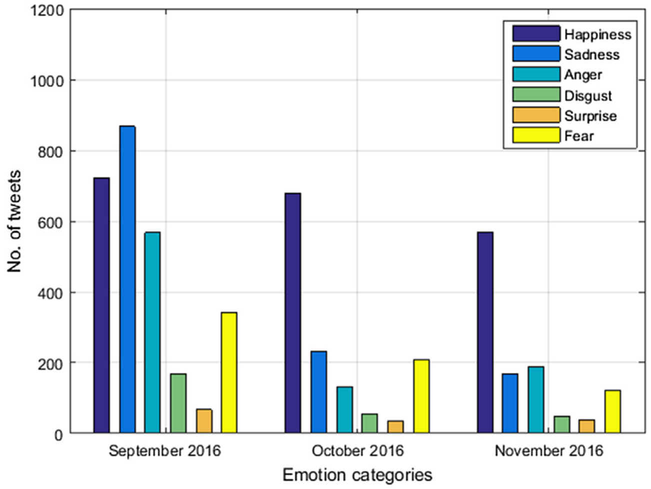

Fig. 4. Sentiment classification by emotions.

worldwide. Tracking public sentiment during the outbreak is a hot research area and prediction based on these sentiments is effective in comparison to survey-based methods [17]. Most researchers have approached the problem of sentiment classification as a kind of text classification.

The current approach for classification of sentiment is based on the principle that the overall sentiment of a text document is the aggregation of the sentiment of the words comprising it. These techniques hence look for the presence of appropriate affect words in a text. This technique either uses a corpus-driven approach to assigning affective orientation or scores to words.

In this particular study, two experiments have been performed. The first experiment finds the text polarity using SentiwordNet [35] and AFINN [36] lexicon dictionary. In the second experiment emotion based classification has been performed on Ekman's emotion theory [37]. The optimistic results regarding the polarity predictive capacity of tweets having contents related to the mosquito-borne diseases are illustrated in Fig. 3.

Emotion classification provides rich sources of information related to public health threats [38]. This study shows whether the public mood can be effectively utilized in early discovery or alarming of such events. For the emotion classification, Jain et al. [38] emotion extraction framework has been used which can also deal with multilingual text data.

The experiments presented in this section clearly showed that emotion-related features gives a better understanding of social media data in comparison to Corpus-based feature and hence, help in better decision-making. Fig. 4 illustrates the sentiment classification by Ekman emotion model [37]. These emotion classes help in identifying public emotions during epidemics and hence help government and health agencies in decision making.

6. Topic modeling framework

The discussions through social media can be used to determine and identifying important attitudes, misinformation, and adverse events related to healthcare. The topic models are a valuable and intelligent tool for understanding unstructured data collected from internet sources [30]. This paper presents a framework for extrac-

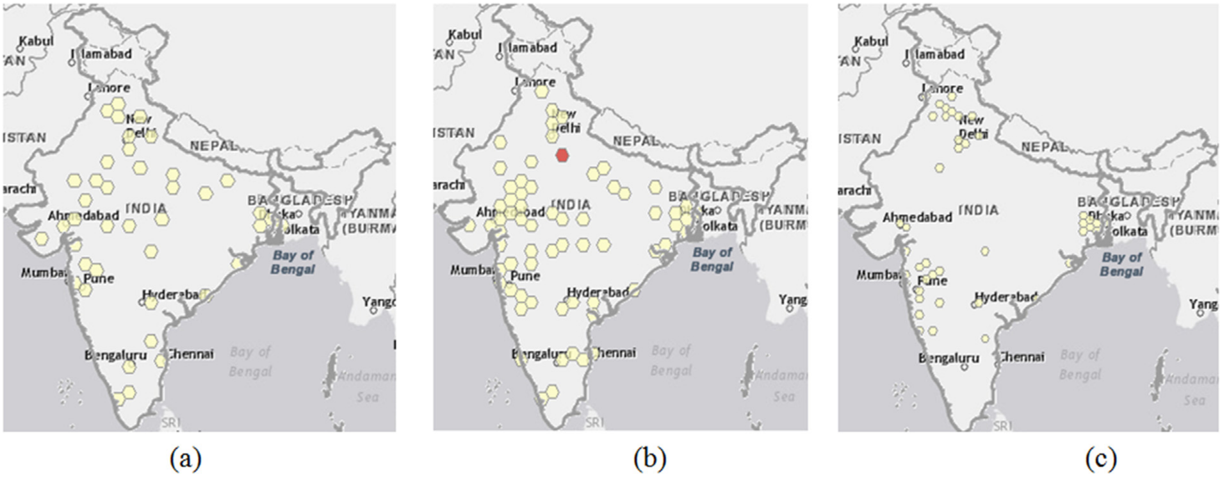

Fig. 5. Geolocation based activity of Twitter users posts related to Dengue and Chikungunya (a) First Phase (September 2016); (b) Second phase (October 2016); (c) Third Phase (November 2016).

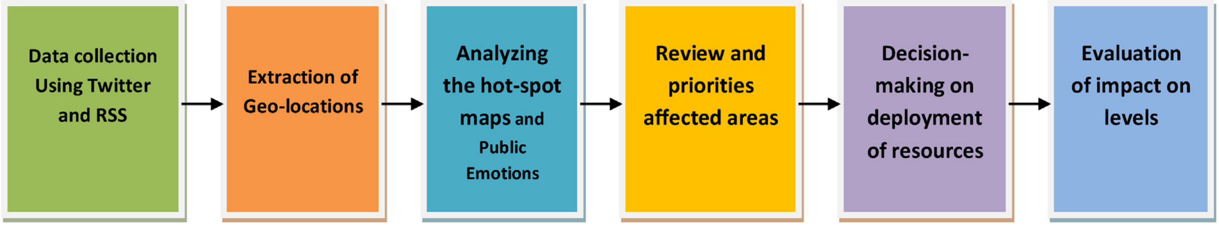
Data collection Using Twitter and RSS
Extraction of Geo-locations
Analyzing the hot-spot maps and Public Emotions
Review and priorities affected areas
Decision-making on deployment of resources
Evaluation of impact on levels

Fig. 6. Intelligent surveillance process model for decision making.

tion of correlations of words presented in the form of opinions. The Latent Dirichlet Allocation (LDA) given by Blei et al. [31], has been used to filter out patterns related to topics and keywords.

The documents are random mixtures of latent topics and a topic is a probability distribution over words. The LDA model helps in mining relevant topics and explores the contents related to public health. In the case of dataset considered here every document (tweet) contains a sequence of words with its multiple probabilities belonging to it. Gibbs sampling has been used for LDA.

The dimensionality (k) of the Dirichlet distribution was known and fixed. The extraction of all the possible combination of topics has been considered as $k = 20$ in the initial stage. The symmetrical Dirichlet priors in the LDA estimation has been considered as $\alpha = 50/K$ and $\beta = 0.01$.

The extraction process gives 20 topics which are and with relevant similar topics. The topics have been classified into three categories: symptoms, prevention, and fear. The topics and words for each relevant topic are extracted manually and divided into three categories which are presented in Table 3.

7. Geo-location extraction

The research has been focused on identifying geo-location for area of intersect in terms of city-level or district level. Another direction is the possibility of predicting at other larger levels of granularity, such as state and specific geographic region. The methodology for proper identification of location has been performed using Twitter and RSS(Rich site summary)feeds of news articles. The fundamental advantage of building up these strategies is two-overlay. To start with, the yield can be utilized to make location-based visualizations and applications. As another illustration, the government health agencies may track drifting conclusions about the proper availability of healthcare crosswise over geologies. Fig. 5 illustrates the approximate region using geo-location extracted by both the methods.

7.1. Geo-location extraction using twitter

The extraction of geo-location of the tweet is one of the crucial steps for identification of correct geographical region. This task has been done by two methods (a) Extracting geo-codes from tweets and; (b) Extracting location based on the content of the tweets. Content-based strategies have additionally been utilized to decide the geo-area of a tweet or, on the other hand to concentrate area data from tweets. The data has been passed to Python library package ‘geograpy’ for extraction of regions and cities. Fig. 5(a–c) illustrated the activity of Twitter users posts related to vector-borne diseases in three different time phases.

7.2. Geo-location extraction using RSS feeds

The architecture for the extraction of geo-locations is based on contents present in RSS feeds of news articles. Generally, RSS feeds contain geo-locations in the form of as name of city, place, state or districts. For extracting geo-location an automatic news article contents extraction system filtered out the main headlines related to Dengue and Chikungunya. These headlines are feed to the Python library ‘geograpy’ for extraction of regions and cities.

8. Proposed predictive model

Social media users are very heterogeneous due to geographic areas. Therefore, the distribution of diseases factors will not have a normal distribution. Generally, the distribution will have a heavy tail since every suffered population will have those extreme situations which need extraordinary attention and precautions. The proposed framework focused on alternative methods of analysis

Table 3
Extracted features words related to mosquito-borne disease.

Symptoms		Prevention		Fear
Hemorrhagic	Fatigue	Mosquito Control	Blood Test	Deaths
Fever	Dizziness	Kapoor	Larva Control	Spread
Back Pain	Cough	Repellents	Empty Tanks	Reported Cases
Headache	Asthma	Tulsi	Fill Holes	Blood
Skin Rash	Abdominal Pain	Goat Milk	No Medications	Terrible
Muscle Pain	Bronchitis	Antibiotic	Homeopathy	Money
Joint Pain	Cold	Protective Clothing	Papita	Treatment
Swelling	WBC	Drain Water	Neem	Rash
Yellow Fever	RBC	Coconut Oil	Onion	Virus
Vomit	Platelets	Fogging	Lemon	Mosquito Attack
Diarrhea	Runny Nose	Insecticide	Ayurvedic	Thunder
Nausea	Itching	Mosquito Net	Turmeric	Kill
Conjunctivitis	Joint Swelling	Empty Coolers	Wear Socks	Outbreak
Pain Eyes	Breathing Trouble	Bed Netting	Leaks	Outburst
Blocked Nose	Respiratory	Close Windows	Full pants	Crash
Pneumonia	Sore Throat	Sleeved Shirts	Avoid Touching	Danger
Bleeding	Boils	Colored Clothing	reduce mosquito	Horror

and visualization of user's opinions that do not depend upon the assumption of normality and historical data. Fig. 6 illustrated the intelligent surveillance process model for decision making.

To identify the affected locations of mosquito-borne disease outbreak **D**, a training window has been defined for an interval of time (September 1, 2016–November 30, 2016). The labeled points have been assigned to the important cities in India using latitude/longitude pairs.

Taking all relevant points, a binary classifier has been trained given in Eq. (4):

$$Pr(Labelp = D|f1(p), f2(p), \ldots, fn(p)) = F(f1(p), f2(p), \ldots, fn(p)) \quad (1)$$

Eq. (1) represents the probability of mosquito-borne disease outbreak D, occurring at a special spatial point *p* which is equal to some logistic function F of the *n* features $f_1(p)$, $f_2(p)$, . . ., $f_n(p)$ [30].

Now taking feature $f_1(p)$, which quantifies the historical mosquito-borne disease density at point *p* .The value has been set for $f_1(p)$has been presented in Equation(5):

$$f_1(p) = k(p, h) = \frac{1}{Ph}\sum_{j=1}^{p} K\left(\frac{\|p - p_j\|}{h}\right) \quad (2)$$

Taking features $f_2(p)$, . . ., $f_n(p)$, which has been derived from tweets in the spatial vicinity of *p*. In Eq. (2):

p represents the point at which a density estimate is needed
h represents the bandwidth parameter
P represents the total number of mosquito-borne disease type T
j indexes a single disease location
K represents a density function

This model uses the combination of the standard kernel density estimate with additional important features using the twitter with RSS feed contents. The model uses latent Dirichlet allocation (LDA) given in Section 6 for identification of coherent topics from collected data set at the particular interval. Fig. 7 presented the Kernel density estimation for tweets that originated within Indian containing words related to mosquito-borne diseases.

Maximum neighborhood has been located by considering the strongest topic Dengue for particular time window with probability 0.32 that contains following symptoms words:{hemorrhagic fever, back pain, skin rash, mosquito repellents, empty tanks muscle pain, headache, joint pain, swelling, yellow fever}

Thus, for any given point *p* falling into this neighborhood, there exists a feature $f_i(p) = 0.32$. The same point *p* is also associated with the other, D − 1 topic probabilities, producing the full set of topic features $\{f_2(p), \ldots, f_i(p) = 0.38, \ldots, fn(p)\}$ for point *p*.

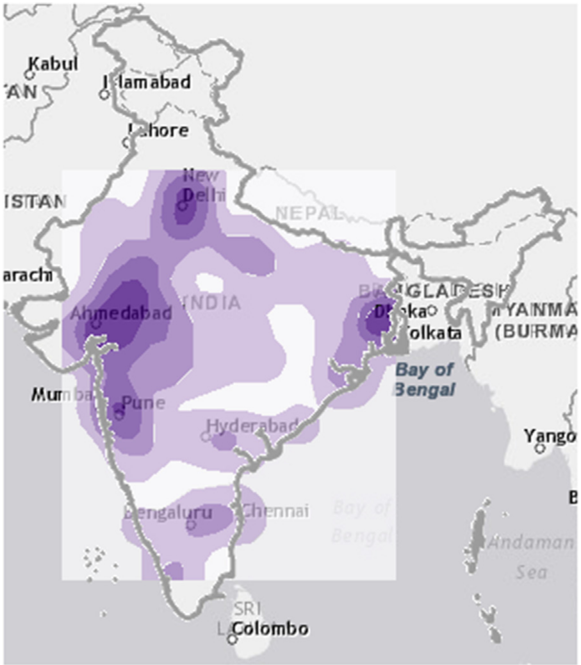

Fig. 7. Predicted surface using Kernel density estimation and Twitter features within Indian.

9. Emergency management model

The optimum utilization of time and resources for an incident which is likely to happen is a measure of the effectiveness of any emergency response service provider system [32]. The quick recovery actions should be taken in time to reduce the loss of life. A new framework has been presented which help government agencies for proper management of time and resources.

It is assumed that health agencies trying to ascertain the best locations for new and old emergency healthcare facility center. The population of interest is all patient suffered from vector-borne diseases in a particular location. Thus the patient distribution should be estimated from a sample derived from the population at large.

Suppose the probability of patient ***p*** ending up at new or existing emergency facility ***h*** is proportional to an exponential function of

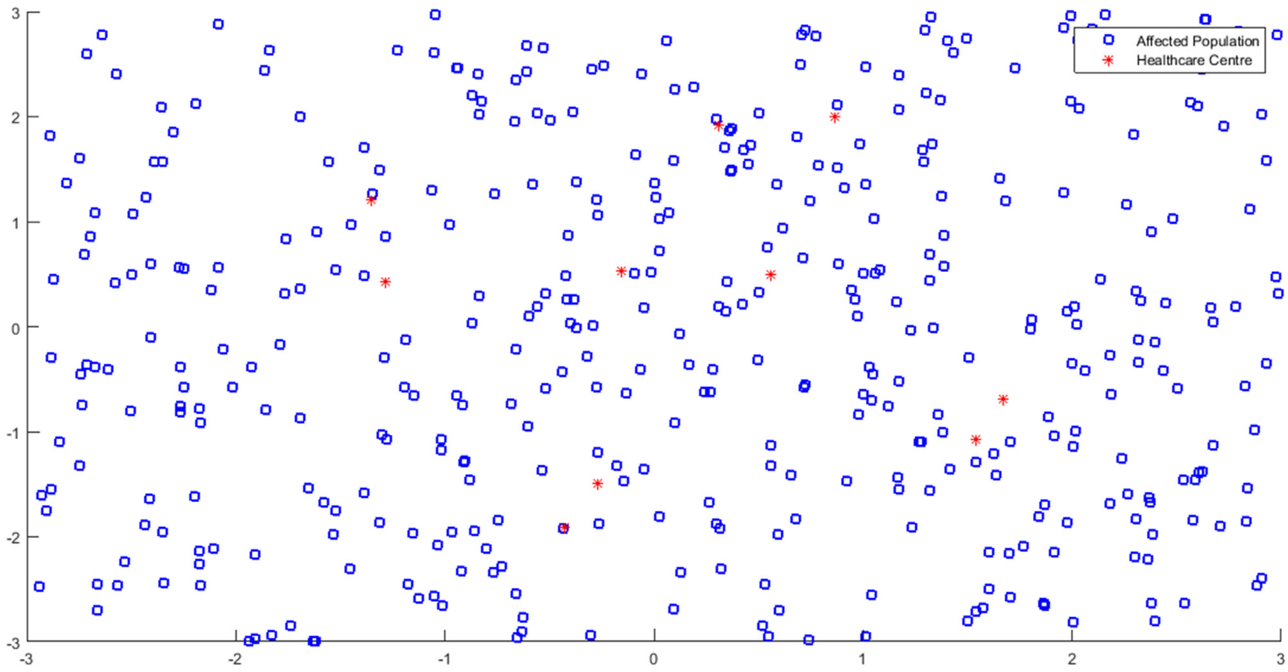

Fig. 8. Scatter plot of random population (400) and sentinel healthcare center (10).

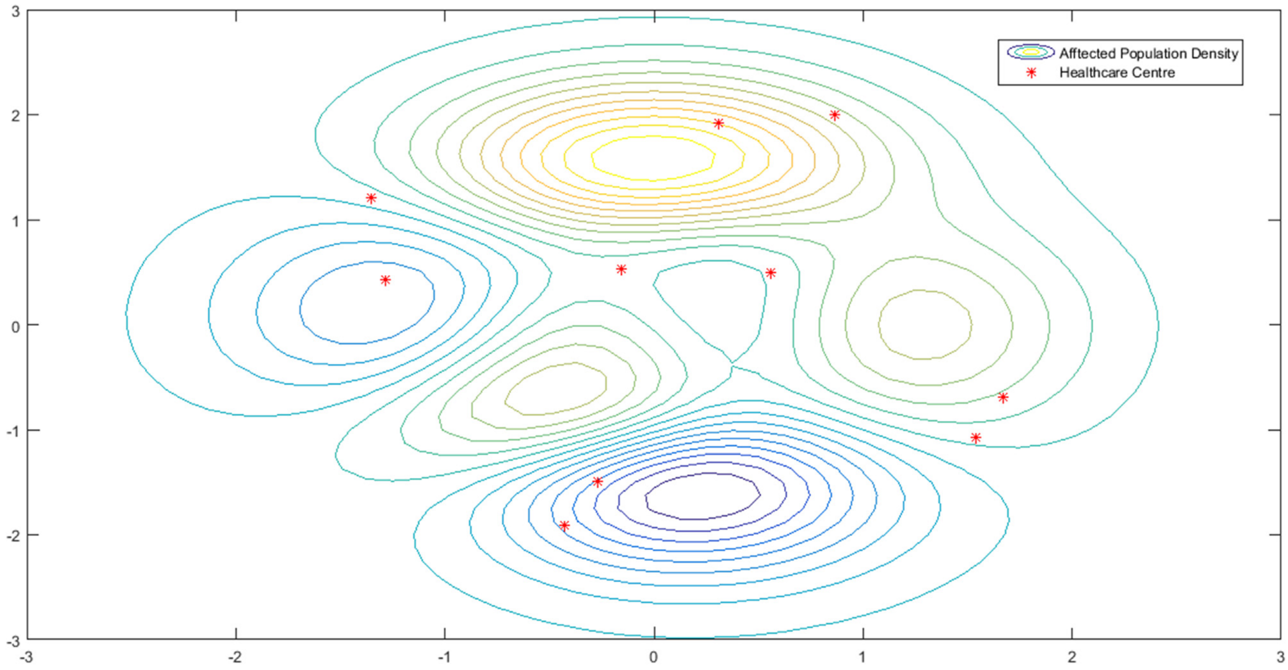

Fig. 9. Contour plot of population density.

the distance between facility **h** and the residence or location of patient ***p***, which results in a Luce formulation of the probability of patient ***p*** arriving at facility ***h***.

$$P_h\left(x_p, y_h\right) = U_{ph}/\sum_{k=1}^{k} U_{pk} \quad (3)$$

Where K is the total number of healthcare centre in the area and

$$U_{ph} = \exp\left(-Cd^2ph\right) \quad (4)$$

Where $\boldsymbol{d_{ph}}$ is the distance between patient ***p*** location and sentinel health center ***j***, ***C*** is a constant representing the current facilities in a particular location and $\boldsymbol{(x_p,y_h)}$ is the location of patient ***p*** location. The value of **C** can easily be calculated from historical records of district hospitals.

For opening or up-gradation of sentinel healthcare center ***h*** it is necessary to find out the density of patients ***g(x,y).*** We assume that each resident of affected area has equal probability of becoming a patient. The additional information can upgrade or downgrade with real-time survey data for geographical locations.

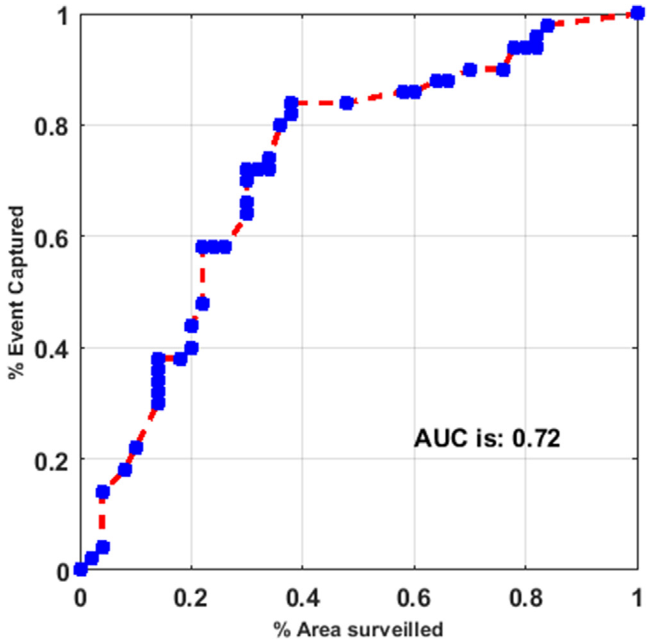

Fig. 10. Surveillance plot showing the number incident of mosquito-borne disease (y-axis) in the x% most threatened area.

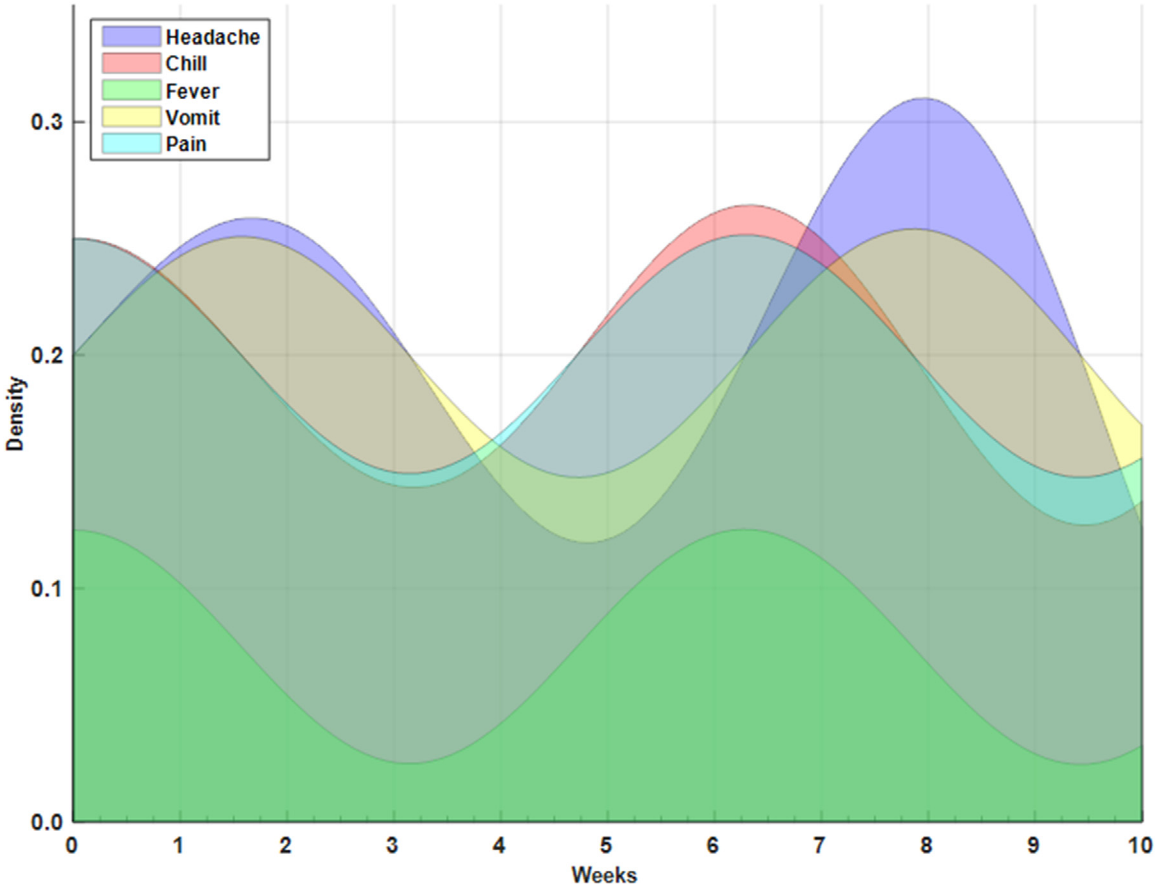

Fig. 11. Density plots of mosquito-borne disease symptoms.

Given patient density ***g(x,y),*** the requirement of healthcare center *(HC$_h$)*with facility his estimated to be***:***

$$HC_h = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)\,P_h(x,y)\,dxdy \qquad (5)$$

This expression takes the existing location into account, because the expression for ***P$_h$(x,y)*** is a function of all competing locations, in addition to the location for facility ***h***. The coordinated which maximize the need for facility ***h*** may be approximated by Taylor series expansion of the above double integration [32].

The method is illustrated on the data from a city which has 10 healthcare centers with random initial population size = 400. Fig. 8 shows a scatter-plot of human location present in a particular geographic region.

It is apparent by the inspection that four of the ten existing healthcare center clustered together and illustrated in Fig. 9.The analysis clear depict that there are clustered which urgently need health care centers and some have to be upgraded.

10. Evaluation and results

For each mosquito-borne disease outbreak type D, Comparative evaluation of the model using only the KDE feature $f_1(p)$ with the model combining $f_1(p)$ with features $f_2(p), \ldots, f_n(p)$ derived from Twitter topics. For identifying the topic probabilities, MALLET tool [33] has been used with configuration with 5000 Gibbs sampling iterations and an optimization interval time window. For calculating the coefficients within the logistic regression mode LibLinear [34] has been used.

For evaluating the model effectiveness, a surveillance plot which measures the percentage of mosquito-borne disease during the prediction window (y-axis) that occurs within the x% most affected area has been plotted is shown in Fig. 10. The better expectation execution is shown by curves that approach the upper-left corner of the plot range or, proportionately, by curves with higher AUC(Area Under the Curve) scores This property makes surveillance plots appropriate for decision makers, who must allocate essential resources (e.g., Healthcare centers, vaccines) across the geographic space.

The health-related conversations on social media have been a goldmine for acquiring information related to disease transmission. Nowadays social media users are extremely open and share their medical problems using multiple social media platforms. An experiment has been conducted to examine the density of mosquito-borne disease symptoms in a time window frame and has been presented in Fig. 11.

These density symptoms help in understanding the diffusion of disease starts to manifest itself. This provides a way to visualize the physical symptoms and geographic reach. Twitter data provide a real-time first source of discussion, suggestions, complaints etc for a particular disease, making it easier to find patterns and new cases. The public sentiments with proper identification of locations or regions provide a clear understanding of where diseases spread. The hope is that Health agencies and epidemiologists will be able to communicate and address the risks in real-time [39].

One of the essential characteristics that drew from this experimental analysis is that Twitter provides real-time information in comparison to traditional research methods involving surveys, panels, and studies and which take months to gather data and finding meaningful inferences. Twitter information was an ideal decision in that it in a flash surfaced practically identical information on a worldwide scale.

11. Advantages of proposed work

The proposed models have been used in multiple data-driven applications which focused on the hidden information contained in a text. An application such as topic-based text categorization, summarization, question answering systems, and information retrieval systems can be improved using proposed method.

Predictive mapping is widely used in developing interfaces which provide appropriate prediction values that help government health agencies in rapid decision making. These models are suitable for improvisation of traditional methods for public health care surveillance.

The extraction of meaningful terms and topics from social media towards events related to healthcare enable researchers and scientist to models better predictive system. The health agencies can take a rapid decision during epidemics and natural calamities.

12. Conclusion and scope of future work

Traditional techniques with social media data offer unique challenges and opportunities for in decision-making in different domains. Traditional methods of surveillance rely on bureaucratic, hierarchical, health-care system which adds a lot of time delays in detection that cannot usually gear up with the speed at which a disease is spread. The proposed framework focused on alternative methods of analysis and visualization of user's opinions that do not depend upon the assumption of normality and historical data. An intelligent surveillance process model for decision making has been proposed through this paper. The classification of the dataset has been performed using machine learning techniques in two phases which provide better results in comparison to other approaches proposed by other authors. A surveillance plot which measures the percentage of the mosquito-borne disease has been presented which clearly indicated that the proposed predictive mapping improves prediction performance.

The work presented in this paper can be pursued further in several domains. One of the tasks is to consider other data resources such as blogs, news articles, forums etc. Data sets containing emoticons, stickers and other images with texts can also be taken into consideration in future.

References

[1] A. Culotta, Towards detecting influenza epidemics by analyzing Twitter messages, in: Proceedings of the First Workshop on Social Media Analytics, ACM, 2010, pp. 115–122.

[2] Vasileios Lampos, Nello Cristianini, Tracking the flu pandemic by monitoring the social web, in: 2nd IAPR Workshop on Cognitive Information Processing (CIP 2010), IEEE Press, 2010, pp. 411–416.

[3] C. St Louis, G. Zorlu, Can Twitter predict disease outbreaks? Brit. Med. J. 344 (2012).

[4] S. Asur, B. Huberman, Predicting the future with social media, in: 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, IEEE, 2010, pp. 492–499.

[5] G. Ifrim, B. Shi, I. Brigadir, Event detection in twitter using aggressive filtering and hierarchical tweet clustering, in: Proceedings of the SNOW 2014 Data Challenge, Seoul, Korea, 2014.

[6] C. Vaccari, A. Valeriani, P. Barberá, R. Bonneau, J.T. Jost, J. Nagler, J.A. Tucker, Political expression and action on social media: exploring the relationship between lower- and higher-threshold political activities among twitter users in Italy, J. Comput.-Mediat. Commun. 20 (2) (2015) 221–239.

[7] Vinay Kumar Jain, Shishir Kumar, Big data analytic using cloud computing, in: 2nd IEEE International Conference on Advances in Computing and Communication Engineering (ICACCE 2015), Dehradun, 2015, pp. 667–672.

[8] C.M. Chew, Pandemics in the Age of Twitter: A Content Analysis Ofthe 2009 h1n1 Outbreak Master's Thesis, University of Toronto, 2010, pp. 15–39.

[9] X. Hu, L. Tang, Huan Liu, Enhancing accessibility of microblogging messages using semantic knowledge, in: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, New York, USA, ACM, 2011, pp. 2465–2468.

[10] Vasileios Lampos, Nello Cristianini, Tracking the flu pandemic by monitoring the social web, in: 2nd IAPR Workshop on Cognitive Information Processing (CIP 2010), IEEE Press, 2010, pp. 411–416.

[11] R. Chunara, J.R. Andrews, J.S. Brownstein, Social and news media enable estimation of epidemiological patterns early inthe 2010 haitian cholera outbreak, Am. J. Trop. Med. Hyg. 86 (1) (2012) 39–45.

[12] E. Aramaki, S. Maskawa, M. Morita, Twitter catches the u: detecting influenza epidemics using Twitter, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2011, pp. 1568–1576.

[13] R. Lowe, T.C. Bailey, D.B. Stephenson, R.J. Graham, C.A.S. Coelho, M. Sá Carvalho, C. Barcellos, Spatio-temporal modelling of climate-sensitive disease risk: towards an early warning system for dengue in Brazil, Comp. Geosci. 37 (2011) 371–381.

[14] V. Racloz, R. Ramsey, S. Tong, W. Hu, Surveillance of dengue fever virus: a review of epidemiological models and early warning systems, PLoS Negl. Trop. Dis. 6 (2012).

[15] D.J. S.I.Hay, G.D.Shanks Rogers, M.F. Myers, R.W. Snow, Malaria early warning in Kenya, Trends Parasitol. 17 (2011) 95–99.
[16] M. Thomson, F. Doblas-Reyes, S. Mason, R. Hagedorn, S. Connor, T. Phindela, A. Morse, T. Palmer, Malaria early warnings based on seasonal climate forecasts from multi-model ensembles, Nature 439 (2006) 576–579.
[17] V.K. Jain, S. Kumar, An effective approach to track levels of influenza-A (H1N1) pandemic in India using twitter, Procedia Comput. Sci. 70 (1) (2015) 801–807.
[18] A. Stewart, E. Diaz, Epidemic intelligence: for the crowd, by the crowd, in: Proceedings of the 12th International Conference on Web Engineering, ICWE'12, Berlin, Heidelberg Springer-Verlag, 2012, pp. 504–505.
[19] T. Bodnar, V.C. Barclay, N. Ram, C.S. Tucker, M. Salathé, On the ground validation of online diagnosis with twitter and medical records, WWW Companion 14 (2014) 651–656.
[20] Jon Parker, Yifang Wei, Andrew Yates, Ophir Frieder, Nazli Goharian, A framework for detecting public health trends with Twitter, Proceeding ASONAM '13 Proceedings Ofthe 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (2013) 556–563.
[21] V. Jain, S. Kumar, A novel approach to track public emotions related to epidemics In multilingual data, in: 2nd International Conference and Youth School Information Technology and Nanotechnology (ITNT 2016), Russia, May, 2016, pp. 883–889.
[22] Philip M. Polgreen, Yiling Chen, David M. Pennock Forrest, D. Nelson, Robert A. Weinstein, Using internet searches for influenza surveillance, Clin. Infect. Dis. 47 (11) (2008) 1443–1448.
[23] M. Keller, M. Blench, C. Tolentino, C. Freifeld, K.D. Mandl, A. Mawudeku, G. Eysenbach, S. Brownstein, Use of unstructured event-based reports for global infectious disease surveillance, Emerg. Infect. Dis. 15 (5) (2009) 689–695.
[24] H.A. Carneiro, E. Mylonakis, Google trends: a web-based tool for real-time surveillance of disease outbreaks, Clin. Infect. Dis. 49 (10) (2009) 1557–1564.
[25] CDC. [https://www.cdc.gov/niosh/topics/outdoor/mosquito-borne/] Date Accessed: 21/12/2015.
[26] AMCA [http://www.mosquito.org/mosquito-borne-diseases].
[27] NVBDCP [http://nvbdcp.gov.in/den-cd.html].
[28] Twitter Developer Page. https://dev.twitter.com/docs/. Date Accessed: 01/01/2015.
[29] Y. Bao, C. Quan, L. Wang, F. Ren, The Role of Pre-Processing in Twitter Sentiment Analysis, ICIC, Taiyuan China, 2014, pp. 615–624.
[30] S. Gerber Matthew, Predicting crime using Twitter and kernel density estimation, Decis. Supp. Syst. 61 (2014) 115–125.
[31] David M. Blei, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[32] Naveen Donthu, T. Rust Roland, Estimating geographic customer densities using kernel density estimation, marketing science, Spring 8 (1989) 191–203.
[33] A. K. McCallum, MALLET: A machine learning for language toolkit, 2002. http://mallet.cs.umass.edu.
[34] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, C.-J. Lin, LIBLIN-EAR: a library for large linear classification, J. Mach. Learn. Res. 9 (2008) 1871–1874.
[35] A. Esuli, F. Sebastiani, SentiWordNet: a publicly available lexical resource for opinion mining, in: Proceedings from International Conference on Language Resources and Evaluation (LREC), Genoa, 2006.
[36] Lars Kai Hansen, Adam Arvidsson, Finn Årup Nielsen, Elanor Colleoni, Michael Etter, Good friends, bad news - affect and virality in twitter, The 2011 International Workshop on Social Computing, Network, and Services (SocialComNet 2011) (2017).
[37] P. Ekman, An argument for basic emotions, Cogn. Emot. 6 (1992) 169–200.
[38] Vinay Kumar Jain, Shishir Kumar, Stevan Fernandes, Extraction of emotions from multilingual text using intelligent text processing and computational linguistics, J. Comput. Sci. (Elsevier) (2017), http://dx.doi.org/10.1016/j.jocs.2017.01.010 http://www.sciencedirect.com/science/article/pii/S1877750317301035.
[39] Elaine, Using Twitter data to study the world's health. [https://blog.twitter.com/official/en_us/a/2015/twitter-data-public-health.html].

Vinay Kumar Jain received his B.E. in Computer Science and Engineering in 2009 from Rajiv Gandhi Proudyogiki Vishwavidyala, Bhopal, India and received his M.Tech in Computer Science and Engineering from Jaypee University of Engineering and Technology, Guna,India in 2012. Now, he is pursuing his Ph.D. degree from Jaypee University of Engineering and Technology, Guna, M.P., India. He has published several papers in peer-reviewed International and Scientific Journals. He is also serving as reviewer for several Science Citation Indexed and Scopus Indexed International Journals.

Shishir Kumar in working as Professor the Department of Computer Science and Engineering at Jaypee University of Engineering and Technology, Guna, M.P., India. He has earned PhD in Computer Science in 2005. He has 14 years of teaching and research experience.