ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

Zika Virus Awareness: A Text Mining Approach.

Kavita S Oza[1]*, VS Kumbhar[1], and RK Kamat[2].

[1]Department of Computer Science, Shivaji University, Kolhapur, Maharashtra, India
[2]Department of Electronics, Shivaji University, Kolhapur, Maharashtra, India

ABSTRACT

Bunches of uproar is made about Zika infection all through the world in the last decade. The present paper reports a text mining based approach to explore the public awareness about this virus in India. Data set for the present investigation is created using keywords which have been used by different search engines. The above mentioned dataset is pre-processed for further analysis. Text mining approach is used to analyze these datasets to find frequently used search keyword. The framework developed is useful for adopting the preventive measures in the public health policy regarding such deadly viruses.

Keywords: Zika virus, text mining, word cloud, Google, Bing, YouTube

Corresponding author

ISSN: 0975-8585

INTRODUCTION

In the last few years, there is a hype which is resulting in generation of big data out of Zika virus research and treatments. Illness caused by a flavivirus is a Zika fever due to mosquitoes was first reported in Uganda at Zika forest in 1947. Slowly it affected the whole Africa and some part of Asia. Societies had overlooked it for a long time until it affected humans. It became issue of concern in 2007, after there were many cases notified as epidemic in Africa. Infection with Zika virus causes trivial disease with fever, malaise, rashes conjunctivitis etc. But these symptoms were developed by few patients and most of the other had no symptoms at all. Death from Zika virus is thought to be infrequent. It has been also reported Zika virus affecting fetuses or infants. There is need for guidance for pregnant women so that they can make a decision about personal protection and pregnancy. World Health Organization (WHO) have started providing regular guidance on risk associated with travel to different Zika virus prone countries and have also called upon all countries to share their data to address Zika virus outbreak[1].

Most of the researchers are working on symptoms, causes and prevention, virus vaccination etc. for Zika virus. Here is attempt to put some light on the social awareness of Zika virus in India. To our knowledge this is the work of its kind and we couldn't find any paper of similar type on web. Text mining approach is used to find the most popular keywords used on the search engine while searching the data related to Zika virus. A small comparison is also made about the online behavior of netizens on different search engines while searching.

Text mining It's an area of data mining where unstructured data is processed to find some useful patterns. This unstructured data may be a word file, customer feedback, social media comments, search texts on search engines etc. Here the challenging part is data is in natural language with ambiguity. Due to ambiguity it is difficult to carry out analysis for mining useful patterns. Now adays there are many text mining software available which can convert these texts into numerical values thus providing a structured form for analysis. Text mining can help in finding out the emotions or sentiments of the person who wrote the comment or drafted the email. It plays a vital role in document classification and clustering. For our work we are using text mining to know online behavior of netizens in India about Zika virus awareness.

Literature review

Zika virus infection is known as a benign infection because of its geographical expansion.A study was carried out to find genetic relationship among Zika viral strains. 32 zika virus isolates were investigated to evaluate the viral spread and its molecular epidemiology [2].The reports till now have indicated that noone with Zika virus infection has been hospitalized. But in some cases Zika virus infection is followed by Guillian-Barre syndrome . A case study related to this has been carried out by [3]. Many researchers are using machine learning approach to mine negative symptoms patterns in electronic health records. As the data is present in electronic form it helps in using different machine learning algorithms like support vector machine, text mining etc to get data in decision making form. [4].

People are still not sure about transmission, health issues and vaccination related to Zika virus. This uncertainty may be due to social media posts and discussions which are sometimes conflicting. There is need for social awareness about zika virus vaccination. There are hindrances to this as there are many websites available on internet which are opposing vaccination scheme. Now a days patients trust more of data on internet than what has been suggested by the doctors. Even for vaccination they surf and browse internet for advice. User post their own views on Zika virus vaccination due to this user generated content through interaction has become universal and followed by many netizens. [5,6,7]

Data collection and preprocessing

Health is one of the important priorities of human beings. As rightly said health is wealth. Now a days people are very cautious about health issues. There are drugs available for almost all the diseases but still there are new chronic disease coming up every day. People need to be aware of these. One way of getting awareness is using internet. Here an attempt is made to know awareness of Zika virus in India amongst netizens.

ISSN: 0975-8585

For this data is collected related to search carried out on search engines like Google and Bing and also on YouTube. A dataset is created containing keywords used by the users while searching on internet about Zika virus.

Table 1: Data set for experiment

Data set details		
Sample Google data set: are zika virus babies retarded are zika virus babies mentally retarded are zika virus mosquitoes in the us are zika virus mosquitoes in florida are zika virus babies brain damage are zika virus babies normal **Preprocessed data:** babies retarded babies mentally retarded mosquitoes us mosquitoes florida babies brain damage babies normal	**Sample Bing dataset:** zika virus in india zika virus symptoms zika virus wiki zika virus ppt zika virus india zika virus vaccine **Preprocessed data:** india symptoms wiki ppt india vaccine	**Sample YouTube dataset:** zika virus documentary zika virus brazil zika virus india zika virus symptoms zika virus in hindi zika virus vaccine **Preprocessed data:** documentary brazil india symptoms hindi vaccine

This data set is preprocessed for analysis. The analysis is carried out to study awareness of people in India about the virus which has not yet reached India. Three text files are created one for Google keywords , second for Bing and third for YouTube. Following table shows the instances in each dataset.

Table 2: size of datasets

Sr. no	Internet tool	No of instances
1	Google	127
2	Bing	144
3	YouTube	224

These instances are processed using Rstudio [8] to know most frequently used keyword while searching for Zika virus on internet.Each data set is preprocessed to remove numerical values if any from the dataset. All the text is converted to lower case. English stop words are also removed. Document stemming is also carried out as part of preprocessing.

RESULTS AND DISCUSSIONS

The pre-processed text is then used to build a document -term matrix. Here documents are represented as rows and terms are represented in columns. Frequent words and associations are taken from this matrix. A word cloud is used to present frequently occurring words in documents. Each document is processed separately and word cloud showing top 20 frequently used keywords. First data set used to create word cloud is of search key words used in Google search engine. Following are the words with frequencies.

can(13) , spread(10), will(10), babies(9), affect(7), cause(5)

ISSN: 0975-8585

Figure 1: word cloud for Google search keywords

Figure 1. shows top 20 frequently occurring keywords in Google searching about Zika virus. Most prominent word which can be seen is can, indicating all the searches like what Zika virus can do , can it be stopped or prevented etc. Next prominent keyword is spread indicating people are more interested in knowing how it spreads. Next concern is related to babies , how babies will be affected by the virus. Following are the concerns related to pregnancy, causes of death its future in India , how does it kill and how far it will stay in India etc.

As per Google search keywords people in India are more worried about the spreading of Zika virus , babies being affected and there is no search related to how to prevent or precautions to be taken, or medicines available.
Second word cloud is constructed using the search keywords used in Bing search engine of Microsoft.
following are the most frequently used keywords in searches.

india (15) hindi(9) virus(9) babi(6) jika(6) new(6)

Analysis of these keyword indicate that people who are using Bing search are from north India as they are searching for results in Hindi Language. They want the details in Hindi language as the second most frequently used keyword is Hindi. Another observation is spelling of Zika as jika.

Figure 2: Word cloud for Bing search keywords

As observed in figure 2, keyword India appears maximum number of times indicating the entire search related to Zika virus are somehow related to India. People are more interested in knowing about Zika virus in Hindi. Here people are also searching about vaccines , symptoms , news on Zika , Zika affected regions etc. As compared to people searching on google , people searching on Bing are more concerned about Zika virus.

As per Bing search keywords, people wants most of the data in Hindi and are more cautious about Zika virus by searching symptoms, affected regions, vaccines for Zika etc.

Third word cloud is constructed using the search keywords of YouTube. YouTube is video search engine. It has large collection of video library. Here searching requires more time and fast internet connection as video list is displayed as part of search result. But we all agree on the fact that we learn more when we see rather than we read. So number of records related to YouTube search is more than Google and Bing as shown in table 1.

Following are most frequently used keywords with their frequencies in YouTube search.

ISSN: 0975-8585

new(11), telugu(6), india(5), babi(4), hindi(4), world(4)

Figure 3: Word cloud for YouTube search

News related to Zika is viewed most of the times as observed in fig. 3. Another observation is telugu keyword indicating people from Andhra Pradesh have viewed more on Zika virus as compared to other states of India. rest of the keywords are like videos on Zika virus, babies affected , spreading of Zika virus, outbreak of Zika virus etc.

As per search carried out on YouTube news related to Zika virus is most prominently watched followed by watching of videos in telugu language on Zika Virus. People are really concerned about health, spreading of Zika virus, babies being affected, videos in Hindi etc.

To analyze the overall online behaviour of netizens in India about Zika virus awareness all these three data sets of Google, Bing and YouTube are combined into one. know the overall online behaviour of people in India towards Zika virus awareness, all the three datasets are combined to form a single corpus and processing is carried out again. Following are the frequently occurring keywords in the search.

india (25) new(17) spread(15) can(13) hindi (13) babi(12)

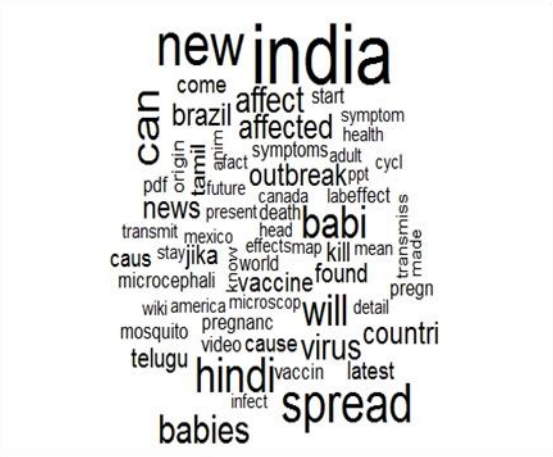

Figure 4: Word cloud for all search keywords

India keyword is used most frequently in search text as seen in figure 4. as the data collected for analysis is from India. This indicates people are searching related to Zika virus in India which shows awareness of it and is good sign. Another prominent keyword is news (news is transformed to new due to stemming) indicating people are relying more on news appearing in paper or TV channels rather than any website. Most of the search is related to spreading of Zika virus not prevention or symptoms of Zika virus. This indicates people are more interested in general knowledge about Zika virus rather than its causes and prevention.

CONCLUSIONS

Overall observation is people in India are not very cautious about Zika virus and searching for it just to remain updated with little knowledge about it . The bad picture is they are not worried about the symptoms, vaccines, medicines, precautions etc related to Zika Virus.

If we try to analyze the different keywords associated with different search engines it shows that people have used YouTube more as compared to other search engines like Google and Bing. This also indicated netizens are not more interested in reading but are interested in audio visual display.

This analysis can be used by the government of India to start new Zika virus awareness programs in India. More and more Zika virus related video can be uploaded on YouTube for masses in Hindi and other regional languages.

REFERENCES

[1] Heymann DL, Hodgson A, Sall AA, Freedman DO, Staples JE, Althabe F, Baruah K, Mahmud G, Kandun N, V as concelos PF, Bino S. Zika virus and micro cephaly: why is this situation a PHEIC?, Lancet 2016; 387(10020):719

[2] Faye O, Freire CCM, Iamarino A, Faye O, de Oliveira JVC, et al. (2014) Molecular Evolution of Zika Virus during Its Emergence in the 20th Century. PLoSNegl Trop Dis 8(1): e2636. doi:10.1371/journal.pntd.0002636

[3] Oehler E, Watrin L, Larre P, Leparc-Goffart I, Lastère S, Valour F, Baudouin L, Mallet HP, Musso D, Ghawche F. Zika virus infection complicated by Guillain-Barrésyndrome – case report, French Polynesia, December 2013. Euro Surveill. 2014;19(9):pii=20720.

[4] Patel, Rashmi et al , Investigation of negative symptoms in schizophrenia with a machine learning text-mining approach, The Lancet , Volume 383 , S16

[5] Mark Dredzea, David A. Bronia towskib,∗, Karen M. Hilyard, Zika vaccine misconceptions: A social media analysis, Vaccine 34 (2016) 3441–3442, Elsevier.

[6] Anna Kata, Anti-vaccine activists, Web 2.0, and the postmodern paradigm–An overview of tactics and tropes used online by the anti-vaccination movement, Vaccine, 30 (25) (2012), pp. 3778–3789.

[7] C.M. Poland, G.A. Poland, Vaccine education spectrum disorder: the importance of incorporating psychological and cognitive models into vaccine education Vaccine, 29 (37) (2011), pp. 6145–6148

[8] R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.