 ***Journal homepage:*** [***http://www.journalijar.com***](http://www.journalijar.com)***INTERNATIONAL JOURNAL***

***OF ADVANCED RESEARCH***

**Research Article**

**A Classification Model to Analyse the Spread and Emerging Trends of the Zika Virus in Twitter**

**\*Rahul Chowdhury1, and Saurabh Thakur2**

**1.** School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, India

**2.** School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, India.

***Manuscript Info Abstract***

***Manuscript History:***

Received: xxxxxxx

Final Accepted: xxxxxxxxxxxxxx

Published Online: xxxxxxxxxxxx

***Key words:***

Zika,

Twitter Analysis,

Text Classification,

Support Vector Machines,

Naïve Bayes Algorithm

,

The recent trend in sharing critical information on social networks such as Twitter has been a motivation for us to propose a classification model that classifies tweets related to Zika and thus enables us to extract helpful insights for the community. In this paper, we try to explain the process of data collection from Twitter, the pre-processing of the data, building a model to fit the data and present some useful predictions and insights that will be helpful in the fight against the Zika virus*.*

*Copy Right, IJAR, 2013,. All rights reserved.*

**Introduction**

The zika virus, is responsible for causing the Zika disease and is primarily carried by the *Aedes* species mosquito. The incubation period of the disease lasts for at most a week and has symptoms such as fever, rashes, headache and conjunctivitis.

Zika virus was declared as a Public Health Emergency of International Concern (PHEIC) by World Health Organization (WHO) on 1st February, 2016. At present, there are no cure such as vaccines or any other form of treatment for this disease and thus makes it a serious global health issue.

Social networking such as Twitter, Facebook etc. have often been treated as useful sources of information for community support on social outbreaks, especially on the global spectrum. Twitter is a popular micro blogging website where users interact socially by posting messages or so called ‘tweets’ on the Twitter platform. Twitter data has been previously been used for various data analysis such as sentiment analysis, event detection etc. and can be easily accessed by the publicly available Twitter API. Twitter is highly popular in mobile application throughout the world and the users can post tweets that can be considered as precise sources of information as they have a 140 character limit. Moreover, there are many verified accounts of reputed people, organizations and communities and thus add more credibility to the tweets.

**Pre-processing of the tweets:** The Twitter Streaming API was used to collect the most recent tweets. The tweets collected by the API are then pre-processed initially to make the later analysis easier. The URL’s, hashtags and user mentions are separated from the text in the original tweet. We also provide an initial analysis of the tweets such as showing graphically countries from where tweets related to Zika are being tweeted the most.

**Building the Classification Model:** After the pre-processing of the collected tweets, we divide our initial data set into training data set and testing data set having 67% and 33% number of tweets respectively. All the tweets in the training data set belong to either of the 3 classes – ‘fight and prevention’, ‘cure’, ‘infected and death’. We then use the Support Vector Machine (SVM) algorithm Naïve Bayes algorithm to train our data and evaluate the accuracy of our methodology using the training data set.

**Comparisons:** The accuracy of the SVM and Naïve Bayes algorithm is compared and then we justify why SVM was chosen as the final classification algorithm for the empirical model.

**Analysing Tweets and Community Support:** After building the intelligent model and determining the accuracy of the empirical model, we have demonstrated how social networks such as Twitter can be used to gather community support about diseases like Zika by analysing the classified tweets.

Few research has been done in building intelligent models for community support on social networking sites and thus our approach demonstrates one such novel method.

**I. Pre-Processing of Tweets**

A Python script was written which with the help of Twitter Streaming API which collects the most recent raw tweets with keywords such as ‘Zika’, ‘Zika Virus’ ,‘Aedes’ in a text file and then each tweet is converted into JSON (JavaScript Object Notation) for easy manipulation and handling of data. Pandas, an open source library for data manipulation in Python is then used to store the data in a data frame with columns such as Twitter ID, created-at, text, favourite-counts etc.

A total of 4751 tweets were collected and after removing the re-tweets, we were left with 1471 unique tweets. The original tweet contains many elements other than the original text such as hashtags, external links, user mentions etc. Thus, for proper analysis of the tweet, from the text we separated 1) stop words such as ‘a’, ‘an’, ‘the’ etc. 2) user mentions 3) hashtags 4) URL’s or external website links 5) special characters such as emoticons. This process of segregation left us with the tweet containing only the main words. A special type of analytical methodology called the word-clouds was then used which when given an array of words, gives us insight into what words have the highest frequency and are important for the analysis. Therefore, world-clouds were generated for main text, hashtags and user-mentions.

The training tweets were then given class label according to the 3 classes – 1) Tweets related to fight and prevention against Zika. 2) Tweets related to cure for Zika. 3) Tweets related to damage caused by the Zika virus, mainly the infected areas and the death caused. Word-clouds were also generated for each of the three classes.

dam was constructed.

**II. Empirical Model**

In this section, we propose a novel system for classifying tweets related to ‘Zika’. The system architecture is shown in Figure 2.1.

Figure 2.1. System Architecture

Tweet Pre-processor

Tweet Scraper

Twitter Server

Streaming API Raw Tweets

Processed Tweets

Classifier Model

Classified Tweets

Prediction and Analysis

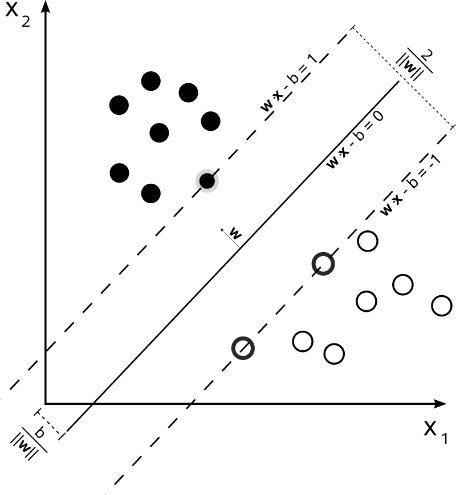
As mentioned, we decided to use the Support Vector Machine (SVM) algorithm and Naïve Bayes Algorithm for our classification as it has been seen earlier that both SVM and Naïve Bayes Algorithm are suitable algorithms for text classification.

The Support Vector Machine (SVM) algorithm is a non-probabilistic binary linear classifier. The model represents data entities as points on a sample coordinate plane in such a way that there is a clear gap between the groups of entities of different classes. The reason SVM work very well for text categorization is that text categorization involves many features (sometimes more than 5000) and SVM handles large feature space.

If there *n* points of the form (x1, y1), … , (xn, yn) where yi is the class for xi  , then it is possible to draw a maximum margin hyperplane between groups having yi =1 and yi =-1 and the hyperplane can be expressed in the following form:

*w.x – b = 1 and w.x – b = -1*

where w is the vector normal to the hyperplane and is shown in Figure 2.2



*Figure 2.2.*

Naïve Bayes’ classifiers are simple probabilistic linear classifiers and is based on the Bayes’ theorem. All Naïve Bayes’ classifiers assume that the features are independent of each other.

If there are n entities in the feature space represented by a vector x = (x1,… ,xn) , then using the Bayes’ theorem, the conditional probability can be expressed as:

*p(Ck | x) = p(Ck) p(x|Ck)*

*p(x)*

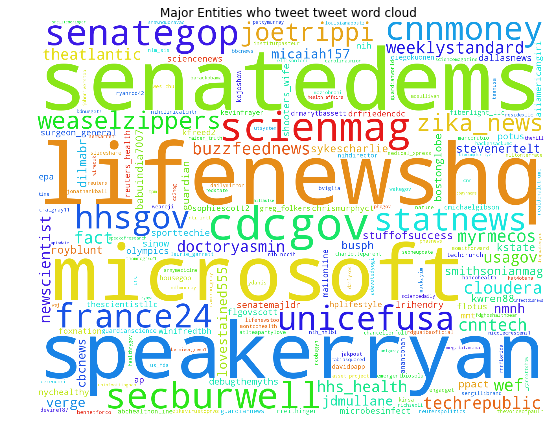
**III. Comparison**

**<Compare SVM and Naïve here>**

**IV. Analysing Tweets and Community Support**

The word clouds generated are depicted in this section. Following are the figures and their description:

Figure 4.1 shows the worldcloud for all the hashtags retreived from the tweets. As depicted, words in bigger font such as ‘zika’ and ‘rio2016’ are the most frequent hashtags.

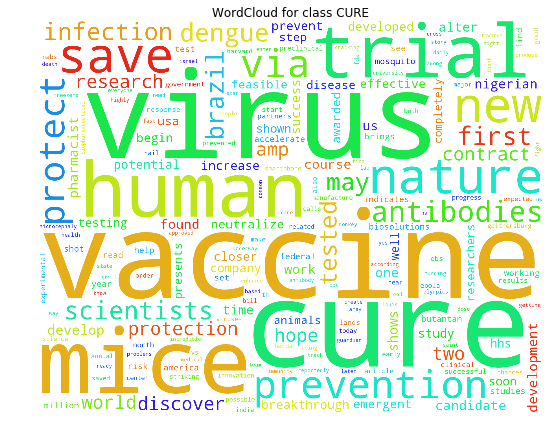
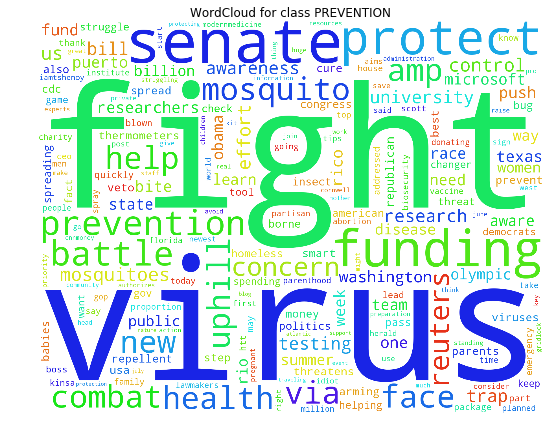


*Figure 4.1*  *Figure 4.2*

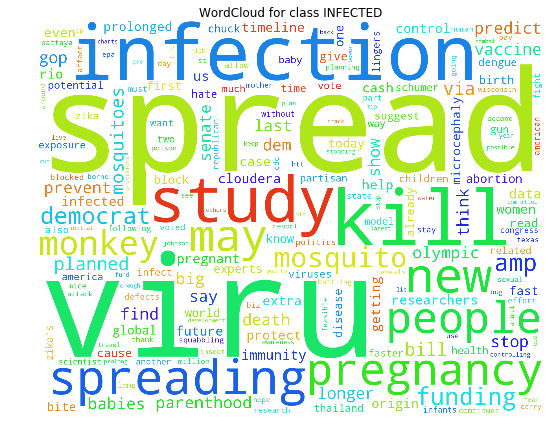
The wordcloud in Figure 4.2 is related to the twitter accounts which have tweeted related to Zika most frequently.

Word clouds were also generated for the 3 classes :

1. Figure 4.3.1 for class ‘fight and prevention’
2. Figure 4.3.2 for class ‘cure’
3. Figure 4.3.3 for class ‘infected and death’



*Figure 4.3.1*  *Figure 4.3.2*



*Figure 4.3.3*

**Result and Discussion**

<conclusion here>

**References**

**Al-Katanani, Y.M., Paula Lopes, F.F. and Hansen, P.J. (2002).** Effect of season and exposure to heat stress on oocyte quality of Holstein cows. J. Daiy Sci. 58: 171-182.

**Butler, W.R. (2001).** Nutritional effects on resumption of ovarian cyclicity and conception rate in postpartum dairy cows. In Diskin M.G. (ed.).

**Dhoble, R.L. and Guplta, S.K. (1981).**Biochemical parameters and response to gonadotrophin administration in anoestrus buffaloes. Indian J. Anim. Sci. 57:47-50.

**Hala, A.A., Abou-Zeina, Hassan, S.G., Sabra, H.A. and Haman, A.M. (2009).** Trials for elevating adverse effect of heat stess in buffaloes with emphasis on metabolic status and fertility. Global Veterinaria 3(1): 51-62.

**Mohammed, F.C., Dhaliwal, G.S. and Sharma, R.K. (1999).** Clinical efficacy of GnRH analogue (Buserelin) and oestradiol benzoate treatment in anoestrus buffaloes. Indian J. Anim. Sci. 69(5): 310-312.

**Nasr, S.M., El Nour, H.H., El-Naggar, A.L., El Said, E.A. and Mahmoud, M.A. (2006).** Clinicopathological studies on dairy cattle suffering from infertility in New Valley governorate. Egyptian J. Comparative. Pathol. Clinc. Patholl. 19: 1-16.

**Reddy, R.K.C., Rao, A.S., Reddy, V.S.C., Yadagiri, B., Sharma, G.P., Rami Reddy, Reddy, M. and Eswa, C. (1994).** Efficacy of certain non hormonal and hormonal drugs on oestrus induction in post partum anoestrus buffaloes. Indian J. Anim. Reprod. 15(2): 127-130.

**Ronchi, B., Stadaioli, G., Verini Suppolizi, A., Bernabuci, U., Lacetera, N. and Accorsi, P.A. (2001).** Infunece of heat stress or feed restriction on plasma progesterone, oestradiol-l7beta, LH, FSH, polactin and cortisol in Holstein heifers. Livestock production Sci. 68: 231-241.

**Thakur, M.S., Jain, P.K. and Rao, M.L.V. (1993).** Induction of oestrus in anoestrus Murrah buffaloes with low doses of Receptal and lutalyse. Inidan J. Anim. Reprod. 17(2): 138.