***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





Streaming API Raw Tweets

Processed Tweets



Classified Tweets

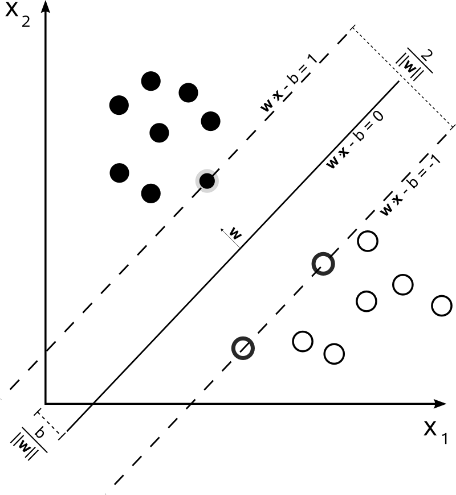
As mentioned, we decided to use the Support Vector Machine (SVM) algorithm and Naïve Bayes Algorithm for our classification as it has been seen earlier that both SVM and Naïve Bayes Algorithm are suitable algorithms for text classification.

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**

Previous research done [16] on classifiers clearly states that SVM and Naïve Bayes’ classifiers give the most promising results when compared to others. But when number of classes are less which in our case is 3 its highly probable that SVM performs better especially due to its non probabilistic approach and its efficiency with high number of features which in our case was around 2804.

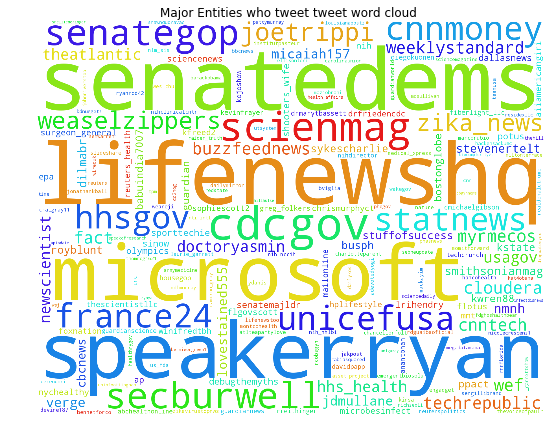
To ensure no overfitting takes place we have used a 5-fold stratified cross validation methodology to get our accuracy score. Table 3.1 presents the mean accuracy score for both SVM and Naïve Bayes’. Clearly SVM offers more scalability due to its linear modeling and is much more accurate.

|  |  |
| --- | --- |
| Classifier Name | 5-fold Stratified Cross Validation Score |
| Multinomial Naïve Bayes’ Classifier | 0.89452 |
| Support Vector Machine (Linear Kernal) | 0.84695 |

**IV. Analysing Tweets Trends 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.

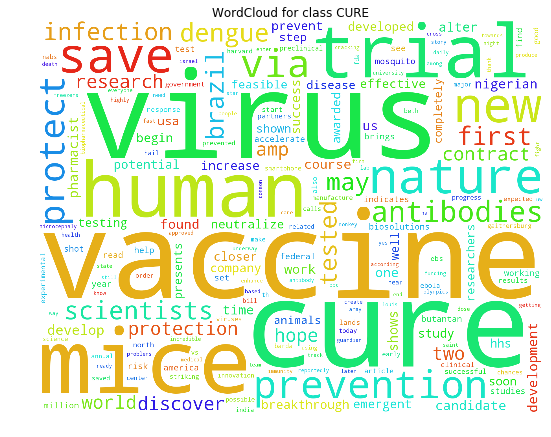
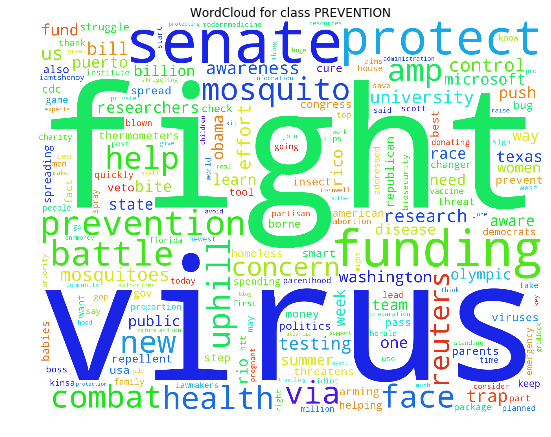


*Figure 4.1*  *Figure 4.2*

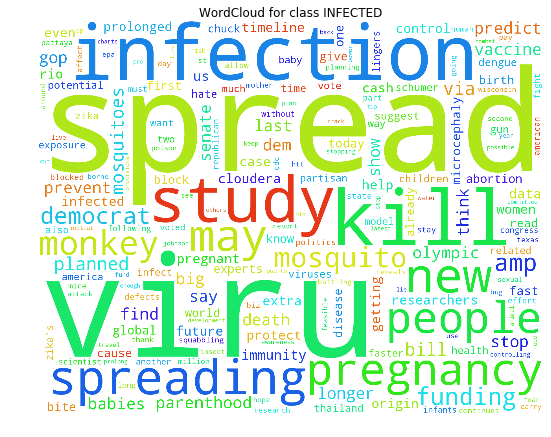
The wordcloud in Figure 4.2 is related to the twitter accounts which have tweeted related to Zika most frequently.

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**

In this paper we present a detailed analytics of most recent tweets collected over a span of two days. We have tried to extract some information out of the tweets which is presented succinctly in the form of word clouds. Finally a model is proposed to scrape and gather most recent tweets using twitter stream API and analyze them in real time and finally classifying them into one of the three broad categories as stated above. We were successful in achieving an accuracy of almost 90 % with Support Vector Machine Classification. This model might find application in community support analysis where one can easily predict what is community more worried about whether its the cure or the spread. Model can also help in making sense of data over time when a huge amount of tweets spanning over several months is available.

For future work we can try observing the behaviour of the model on a relatively larger dataset spanning over a period of 3 to 4 months. A time based analysis and geospatial analysis are some major branches of development for this model. Such model can easily be tweaked to later on analysing various other events and not just epidemics like Zika.

**References:**

[1] http://www.who.int/mediacentre/factsheets/zika/en/

[2] Sakaki, Takeshi, Makoto Okazaki, and Yutaka Matsuo. "Earthquake shakes Twitter users: real-time event detection by social sensors." *Proceedings of the 19th international conference on World wide web*. ACM, 2010.

[3] Pak, Alexander, and Patrick Paroubek. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining." *LREc*. Vol. 10. 2010.

[4] Lerman, Kristina, and Rumi Ghosh. "Information contagion: An empirical study of the spread of news on Digg and Twitter social networks." *ICWSM* 10 (2010): 90-97.

[5] https://dev.twitter.com/streaming/overview

[6] Joachims, Thorsten. "Text categorization with support vector machines: Learning with many relevant features." *European conference on machine learning*. Springer Berlin Heidelberg, 1998.

[7] McCallum, Andrew, and Kamal Nigam. "A comparison of event models for naive bayes text classification." *AAAI-98 workshop on learning for text categorization*. Vol. 752. 1998.

[8] https://www.python.org/

[9] http://www.json.org/

[10] http://pandas.pydata.org/

[11] https://amueller.github.io/word\_cloud/

[12] Sebastiani, Fabrizio. "Machine learning in automated text categorization."*ACM computing surveys (CSUR)* 34.1 (2002): 1-47.

[13] Tong, Simon, and Daphne Koller. "Support vector machine active learning with applications to text classification." *Journal of machine learning research*2.Nov (2001): 45-66.

[14] Shmilovici, Armin. "Support vector machines." *Data Mining and Knowledge Discovery Handbook*. Springer US, 2005. 257-276.

[15] El Kourdi, Mohamed, Amine Bensaid, and Tajje-eddine Rachidi. "Automatic Arabic document categorization based on the Naïve Bayes algorithm."*Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages*. Association for Computational Linguistics, 2004.

[16] Hassan, Sundus, Muhammad Rafi, and Muhammad Shahid Shaikh. "Comparing SVM and naive bayes classifiers for text categorization with Wikitology as knowledge enrichment." *Multitopic Conference (INMIC), 2011 IEEE 14th International*. IEEE, 2011.