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

B.K. Tripathy, Saurabh Thakur and Rahul Chowdhury

Abstract The Zika disease is a 2015–16 virus epidemic and continues to be a global health issue. 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 into the community. In this paper, we try to explain the process of data collection from Twitter, the preprocessing of the data, building a model to fit the data, comparing the accuracy of support vector machines and Naïve Bayes algorithm for text classification and state the reason for the superiority of support vector machine over Naïve Bayes algorithm. Useful analytical tools such as word clouds are also presented in this research work to provide a more sophisticated method to retrieve community support from social networks such as Twitter.

Keywords Zika · Twitter analysis · Twitter classification · Support vector machines · Naïve Bayes algorithm

1 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 February 1, 2016. At present,

B.K. Tripathy (✉) · S. Thakur · R. Chowdhury
School of Computing Science and Engineering, VIT University, Vellore, India
e-mail: tripathybk@vit.ac.in

S. Thakur
e-mail: saurabh.chandrakantthakur2013@vit.ac.in

R. Chowdhury
e-mail: rahul.chowdhury2013@vit.ac.in

H.S. Behera and D.P. Mohapatra (eds.), *Computational Intelligence in Data Mining*, Advances in Intelligent Systems and Computing 556,
DOI 10.1007/978-981-10-3874-7_61

there are no cures 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 and Facebook has often been treated as useful sources of information for community support on social outbreaks, especially on the global spectrum [9]. Twitter is a popular microblogging Web site where users interact socially by posting messages or the so-called ‘tweets’ on the Twitter platform. Twitter data have previously been used for various data analysis such as sentiment analysis [3, 8] and event detection and can be easily accessed by the publicly available Twitter API (application program interface). 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 [4]. Moreover, there are many verified accounts of reputed people, organizations, and communities and thus add more credibility to the tweets.

Preprocessing of the tweets: The Twitter Streaming API was used to collect the most recent tweets. The tweets collected by the API are then preprocessed initially to make the later analysis easier. The URLs, 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 the countries from where tweets related to Zika are being tweeted the most.

2 Preprocessing of Tweets

A Python script was written with the help of Twitter Streaming API which collects the most recent raw tweets with keywords such as ‘Zika,’ ‘Zika virus,’ and ‘Aedes’—the species name of the mosquito causing Zika 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, and favorite-counts.

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, and user mentions. Thus, for proper analysis of the tweet, from the text we separated (1) stop words such as ‘a’, ‘an,’ and ‘the’; (2) user mentions; (3) hashtags; (4) URLs or external Web site links; and (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, word clouds were generated for main text, hashtags, and user mentions.

The training tweets were then given class label according to the three classes—(1) tweets related to fight and prevention against Zika; (2) tweets related to cure for Zika; and (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.

3 Empirical Model

In this section, we propose a novel system for classifying tweets related to 'Zika.' The system architecture is shown in Fig. 1.

Building the Classification Model: After the preprocessing 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 any of the three classes—'fight and prevention,' 'cure,' and 'infected and death' since each class has separate set of keywords, hence preventing overlapping of the classes.

As previous research [8, 9] suggests the efficiency of support vector machine (SVM) algorithm [1, 6] and Naïve Bayes algorithm [7] for text classification, we use them 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.

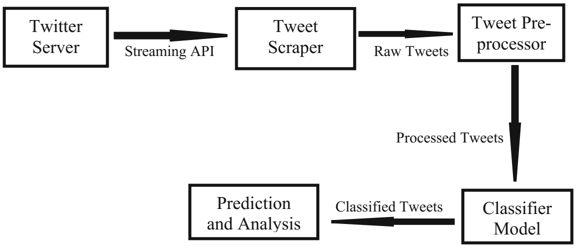
Analyzing Tweets and Community Support: After building the intelligent model and determining the accuracy of the empirical model, we have demonstrated how social networks such as Twitter can be used to gather community support about diseases like Zika by analyzing the classified tweets.

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.

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 [10, 12].

The support vector machine (SVM) algorithm [11] 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.

Fig. 1 Model architecture



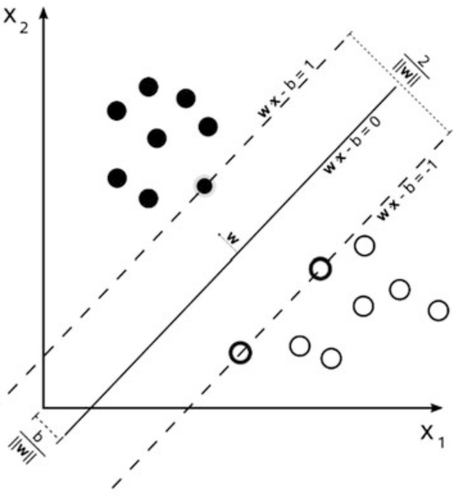


Fig. 2 Maximum margin hyperplane and margins for an SVM from 2 classes

If there are *n* points in the form $(x_1, y_1)\ldots(x_n, y_n)$ where y_i is the class for x_i, then it is possible to draw a maximum margin hyperplane between groups having $y_i = 1$ and $y_i = -1$ and the hyperplane can be expressed in the following form:

$$w.x - b = 1 \text{ and } w.x - b = -1 \; w.x - b = 1 \text{ and } w.x - b = -1 \qquad (1)$$

where w is the vector normal to the hyperplane and is shown in Fig. 2.

Naïve Bayes' classifiers [2] are simple probabilistic linear classifiers and are based on 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 $x = (x_1,\ldots,x_n)$, class C_k then using Bayes' theorem, the conditional probability can be expressed as follows:

$$p(C_k|x) = \frac{p(C_k)p(x|C_k)}{p(x)} \qquad (2)$$

4 Comparison

Previous research done [3, 5] 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, it is 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. Features are nothing but the most commonly occurring words in the corpus of all the tweets.

To ensure no overfitting takes place, we have used a fivefold stratified cross-validation methodology to get our accuracy score. The data were divided into 5 subsamples maintaining the ratio of tweets belonging to each class. For accuracy measures, we used simple classification accuracy metric which shows the fraction of correct predictions over total predictions. ŷ in Eq. 3 is predicted target variable and y is actual target variable.

$$\text{accuracy}(\mathrm{y}, \hat{\mathrm{y}}) = \frac{1}{\eta_{\text{samples}}} \sum_{i=0}^{\eta_{\text{samples}}-1} 1(\hat{\mathrm{y}}_i = \mathrm{y}_i) \tag{3}$$

Table 1 presents the mean accuracy score for both SVM and Naïve Bayes' classifier. Clearly, SVM offers more scalability due to its linear modeling approach and is much more accurate.

5 Analyzing Tweets and Community Support

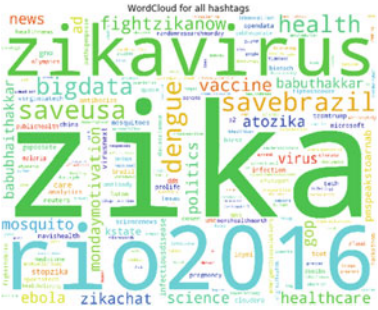
The word clouds generated are depicted in this section. Following are the figures and their description:

Figure 3 shows the word cloud for all the hashtags retrieved from the tweets. As depicted, words in bigger font such as 'Zika' and 'rio2016' are the most frequent hashtags.

Table 1 Table depicting the comparison between Naïve Bayes' classifier and support vector machine

Classifier name	Fivefold stratified crossvalidation score
Multinomial Naïve Bayes' classifier	0.84695
Support vector machine (linear kernel)	0.89452

Fig. 3 Hashtag word cloud



The word cloud in Fig. 4 is related to the twitter accounts which have tweeted related to Zika most frequently.

Word clouds were also generated for the three classes:

(1) Figure 5 for class ‘fight and prevention,’
(2) Figure 6 for class ‘cure,’ and
(3) Figure 7 for class ‘infected and death.’

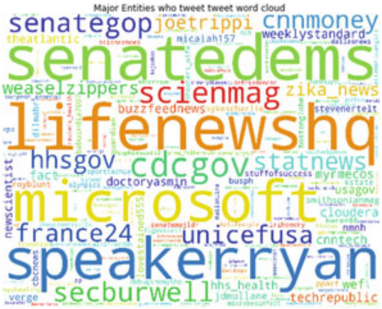


Fig. 4 Twitter mention word cloud

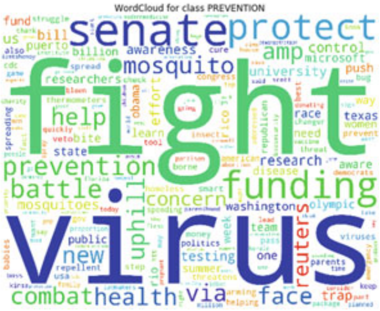


Fig. 5 Fight and prevention

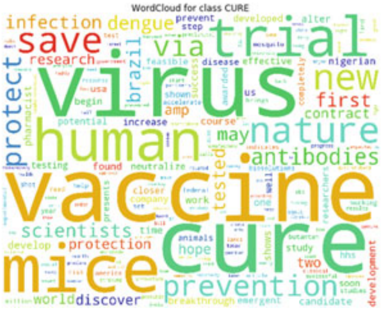


Fig. 6 Cure

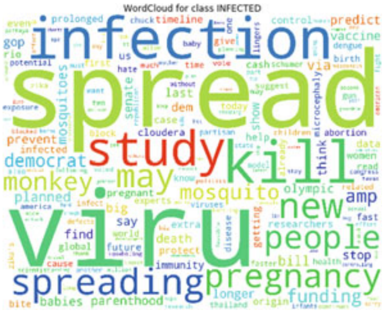


Fig. 7 Infected and death

6 Conclusion

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 that 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. According to the analysis, it was inferred that in the tweets gathered 36.50%, 24.94%, and 38.54% of tweets belonged to ‘fight and prevention,’ ‘cure,’ and ‘infected and death,’ respectively. These values also provide a statistical evidence of social community support and awareness available for Zika presently.

7 Applications and Future Work

This model might find application in community support analysis where one can easily predict what is community more worried about whether it is 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 behavior of the model on a relatively larger data set spanning over a period of 3–4 months. A time-based analysis and geospatial analysis are some major branches of development for this model. It would really be interesting to observe how the community support changes over time. Such model can easily be tweaked to later on analyzing various other events and not just epidemics like Zika.

References

1. Cristianini, Nello, and John Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.
2. 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.
3. 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.
4. Joachims, Thorsten. “Text categorization with support vector machines: Learning with many relevant features.” *European conference on machine learning*. Springer Berlin Heidelberg, 1998.
5. Khan, Aamera ZH, Mohammad Atique, and V. M. Thakare. “Combining lexicon-based and learning-based methods for Twitter sentiment analysis.” *International Journal of Electronics, Communication and Soft Computing Science & Engineering (IJECSCSE)* (2015): 89.
6. 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.
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. Pak, Alexander, and Patrick Paroubek. “Twitter as a Corpus for Sentiment Analysis and Opinion Mining.” *LREc*. Vol. 10. 2010.
9. 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.
10. Sebastiani, Fabrizio. “Machine learning in automated text categorization.” *ACM computing surveys (CSUR)* 34.1 (2002): 1–47.
11. Shmilovici, Armin. “Support vector machines.” *Data Mining and Knowledge Discovery Handbook*. Springer US, 2005. 257–276.
12. 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.