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

*ABSTRACT*

*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.*

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 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 tweets collected by the Twitter Streaming API are 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 3000 collected tweets, we divide our initial data set into training data set and testing data set having 55% and 45% 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 Naïve Bayes algorithm to train our data and evaluate the accuracy of our methodology using the training data set.

**Analysing Tweets and Community Support:** After building the intelligent 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.

1. PRE-PROCESSING OF TWEETS

We wrote a Python script which with the help of Twitter Streaming API which collects 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.

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.