1. EMPIRICAL MODEL

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

Figure 1. System Architecture

Tweet Pre-processor

Tweet Scraper

Streaming API Raw Tweets

Twitter Server

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

Naïve Bayes classifiers are linear classifiers and is based on the Bayes’ theorem.