Journal of Korea Multimedia Society Vol. 22, No. 5, May 2019(pp. 588-600)
https://doi.org/10.9717/kmms.2019.22.5.588

Malaria Epidemic Prediction Model by Using Twitter Data and Precipitation Volume in Nigeria

Nduwayezu Maurice†, Satyabrata Aicha††, Han Suk Young†††, Kim Jung Eon††††, Kim Hoon†††††, Park Junseok††††††, Hwang Won-Joo†††††††

ABSTRACT

Each year Malaria affects over 200 million people worldwide. Particularly, African continent is highly hit by this disease. According to many researches, this continent is ideal for Anopheles mosquitoes which transmit Malaria parasites to thrive. Rainfall volume is one of the major factor favoring the development of these Anopheles in the tropical Sub-Sahara Africa (SSA). However, the surveillance, monitoring and reporting of this epidemic is still poor and bureaucratic only. In our paper, we proposed a method to fast monitor and report Malaria instances by using Social Network Systems (SNS) and precipitation volume in Nigeria. We used Twitter search Application Programming Interface (API) to live-stream Twitter messages mentioning Malaria, preprocessed those Tweets and classified them into Malaria cases in Nigeria by using Support Vector Machine (SVM) classification algorithm and compared those Malaria cases with average precipitation volume. The comparison yielded a correlation of 0.75 between Malaria cases recorded by using Twitter and average precipitations in Nigeria. To ensure the certainty of our classification algorithm, we used an oversampling technique and eliminated the imbalance in our training Tweets.

Key words: Classification Algorithms, Malaria and Precipitation, Pearson Correlation Coefficient, and Twitter Data.

1. INTRODUCTION

Malaria disease monitoring and reporting are still poor and bureaucratic despite the fact that it still poses a severe threat to humanity. Each year more than 200 million people are affected by this disease worldwide particularly SSA. According to the 2018 World Health Organization (WHO) reports released on 2018 November 16 on the year 2017 Malaria cases and deaths [1], 219 millions of Malaria cases occurred worldwide and an estimated 435,000 Malaria deaths were counted.

African regions carry a disproportionately high share of this global Malaria burden. In that 2017 year only, those regions were home to 92% of all worldwide Malaria cases and 93% of Malaria deaths. That report also stated that Nigeria is the first country out of five which accounts for nearly

※ Corresponding Author : Won-Joo Hwang, Address: 197 Injero, Gimhae, Gyeongnam 50834, South Korea, TEL : +82-55-320-3847, FAX : +82-55-322-6275, E-mail : ichwang@inje.ac.kr

Receipt date : Dec. 19, 2018, Revision date : Apr. 24, 2019
Approval date : Apr. 26, 2019

† Dept. of Information and Communication Systems, Inje University (E-mail: nduwayezumaurice@gmail.com)
†† Institute of Digital Anti-Aging Healthcare (IDA), Inje University (E-mail: satyabrataaich@gmail.com)
††† Institute of Digital Anti-Aging Healthcare (IDA), Inje University (E-mail: thewayceo@inje.ac.kr
†††† Dept. of Emergency Medicine, Inje University, Ilsan Paik Hospital
(E-mail: jekim1229@naver.com)
††††† Dept. of Emergency Medicine, Inje University, Ilsan Paik Hospital (E-mail: megali@hanmail.net)
†††††† Dept. of Emergency Medecine, Inje University, Ilsan Paik Hospital (E-mail : edpjs@inje.ac.kr)
††††††† Dept. of Electronics, Telecommunications, Mechanical & Automotive Engineering, Inje University (E-mail: ichwang@inje.ac.kr)

half of all Malaria cases worldwide therefore being the first country highly hit by Malaria with 25% of the total Malaria cases worldwide. The WHO also recommends that effective surveillance of Malaria cases and deaths is essential for identifying the areas or population groups that are most affected by Malaria, and for targeting resources for maximum impact.

One of the attributes of a strong surveillance system according to WHO is the high level of Malaria cases detection.

However, Malaria monitoring and evaluation in most of the endemic countries of SSA has been mainly based on periodic national household surveys [2]. Despite the fact that these methods are advantageous in their adequacy to capture the underlying variation in the sampled population and a flexibility of data collection instruments which can accommodate a number of questions on a variety of topics, the followings are some of their disadvantages. They are expensive, time consuming and labor intensive. If the measures of eliminating this pandemic continues to rely only on those methods, the eradication will remain a dream while the rate at which Malaria takes lives is increasing. In addition, these monitoring would not facilitate researchers to find ways to eradicate this disease because there continue to be a lack of sufficient and real-time Malaria data as the basis for further researches.

In the same WHO Malaria report of 2018, responsible ministries of endemic SSA countries are recommended to collect reports from their health facilities through health information systems in order to improve Malaria surveillance systems. However, there is still a common delay between the actual outbreak and the perception of those ministries. Consequently, the measures by those responsible ministries to stop a spread or react on the outbreak might not have effect because of that delay. Moreover the dominance of this endemic is in the part of the world where there are also economical and financial problems which might impede computerization of those health information systems. Therefore a cheaper, frequent, rapid and whole population covering Malaria surveillance and monitoring systems are indispensably needed. There is also a need to even not only rely on ground data to draw measures and planning about Malaria but also use other indicators like climatology to support the surveillance, reporting and warning population about this endemic.

Fortunately, Social media present a source of huge and rich information data covering different subject areas. These SNSs are increasingly being used as source of data in the monitoring [3] and reporting of many other diseases like Influenza and Ebola. Typically, Twitter as one of the micro-blogging SNS, can provide a robust platform for public health practitioners to detect incidences and manage intervention measures in Malaria prone areas [4]. Also according to a number of researches, the spreading of Malaria pandemic is highly linked to the precipitation volume especially in the tropical seasons therefore precipitation can also be used as another important warning factor to the potential occurrence of Malaria.

Since its creation in 2006, Twitter has been growing steadily and an average of over 200 million messages are currently posted every day. As a result, incredible amounts of information on an immense variety of domains are now available through their service. The Twitter data structure is compact so it forces users to post short comments. As a result, users tend to post short messages in the sense of snapshots moods and feelings as well as for up-to-date events and current situation commenting [5], [6]. This is the essence to why Twitter messages are increasingly used by health care researchers and professionals for surveillance and reporting of many diseases spread.

Malaria disease is caused by the Plasmodium genus that is transmitted between humans by

Anopheles mosquitoes [7]. The Anopheles responsible for transmitting the Plasmodium genus is very sensitive to the environment conditions. Rainfall volume is one of the most frequently reported determinants of Malaria transmission by these Anopheles mosquitoes [8]. This meteorological factor affects the incubation, developmental and survival stages of mosquitoes, and Plasmodium within the mosquito though their effects are delayed to about two weeks for Malaria symptoms to start manifesting [9]. Rainfall plays a significant role in provision of the breeding sites of the Anopheles [7].

There have been some other interesting works that used SNS to study about Malaria and other mosquito born diseases. However, none have directly reported instances of Malaria using these SNS. In addition, they rely only on previous Malaria instances to define or validate their new methods of monitoring this disease. As far as we know no research has been conducted leveraging Twitter to directly estimate current number of Malaria instances in a particular area and comparing with live-precipitation volume in that area. The contribution of our work to the monitoring and reporting of Malaria diseases is summarized in two points as follows;

- We leveraged Twitter in order to fast and real-timely monitor and report on Malaria instances in Nigeria and,
- We studied the relationship between those Malaria instances and the precipitation volume in the same country as a way to both validate our Twitter data and assess the usefulness of precipitation as another important warning factor of Malaria potential occurrence.

2. RELATED WORKS

In the last few decades, research articles interfacing illness diseases connected with web-based social networking have expanded in number because of the expansion in accessibility of real-time data from different geo-locations. In [4], James and Hassan proposed a conceptual Malaria Surveillance System that leverages SNS with a view to enhancing decision making by public health professionals. Their proposed system is composed of a data collection module which uses Twitter search interface API to extract message containing key words like Malaria, outbreak and intervention messages. Their conceptual system uses WHO Malaria data of year 2004-2014 as a basis of evaluation. Interestingly, this was the first ever work to have leveraged Twitter, as they themselves declared, in order to mine the data about Malaria disease. However, they module does not directly estimate the instances of Malaria and still use the past Twitter data and past Center for Diseases Control (CDC) data. A live module that monitor Malaria Tweets is needed and also validating these social media data with other facts like rail fall, not only CDC data, is also helpful for improved and extended Malaria monitoring and reporting strategies. This is where we began and propose a model that directly estimates the malaria instances using Twitter and relates those instances with the precipitation volume in the same period.

A research about tropical diseases mapping in Indonesia in [10] used Twitter Representational State Transfer (REST) API and crawled Tweets related to the diseases dominant in the tropical area of our globe mainly Malaria, Dengue and Avian flu. They used 5-folds Bernoulli Naive Bayes to classify the Tweets into tropical diseases related or not disease and then mapped those diseases tweets to different regions on a map to signal which disease is dominant in a particular region in Indonesia. Their work is interesting as it can help health care professionals allocate exact resources to combat a particular disease in an exact area. Nevertheless, it still does not provide directly real-time estimates on Malaria disease itself. [11] is also another work that leveraged SNSs for the surveillance of diseases. This research studied the prediction of Flu trends using Twitter data. Basing on the data they collected during 2009 and 2010 they found that the volume of Flu related Tweets is highly correlated to the volume of Influenza Like Illness (ILI) cases collected by CDC. Then they applied regressive model on the historic CDC data to forecast the future ILI cases and found that Tweets improved their model of prediction. They concluded that Twitter data provides real-time assessment of ILI activity. This work among others exploit Twitter and retrieved useful insights helpful for diseases monitoring. Unfortunately few of them have covered Malaria so far.

Many researchers have published a strong relationship between Malaria Anopheles development and rainfall. A research by A. Ayanlade et.al. in [12] is quite comparable to ours. In their research they studied the relation between intra-annual climate variability and Malaria transmission in Nigeria. They results show that Malaria is a function of, mainly rainfall, and that temperature or humidity influences the longevity of the Anopheles breedings. They showed monthly climate conditions suitable for Malaria and those conditions were calculated on the basis of the percentage of monthly rainfall, temperature, humidity and Malaria occurrences in different zones of the country of Nigeria. K. Olayemi et al. in [13] studied the climate of North Central Nigeria along with the relation of that climate with Malaria. Their results indicated that north central Nigeria is clearly seasonal and that seasonality influences Malaria transmission. They stated that rainy season provides optimal condition for breeding, human biting activity and survival of Anopheles mosquitoes thus, significantly enhancing the intensity of Malaria transmission during that period. Also A. B. Adigun et al. in [14] and Gbenga J. Abiodum et al. in [9] studied this relationship. The advantages of these three works on Malaria and rain volume relationship is that they proved precipitation as another metric which can be used to predict Malaria disease occurrences. However, most of them are expensive and time consuming. They used not real-time data as the SNS that we proposed which is faster and economically affordable.

In our work we proved that real-time Malaria monitoring model using Twitter can also be used as a reliable and robust system to estimate Malaria cases by comparing with the real-time precipitation volume. We achieved it by applying Pearson Correlation Coefficient on Malaria instances from Twitter and average precipitation calculated over the sum of daily rain in 24 main areas in Nigeria.

3. METHODOLOGY

In our research we streamed live Twitter messages mentioning Malaria. We preprocessed those Tweets and classified them into two different classes namely Malaria cases related Tweets and Malaria cases unrelated Tweets by using SVM classifier. Then we calculated the relationship between those Malaria cases with the precipitation volume in Nigeria. The system model of our work is shown in Fig. 1.

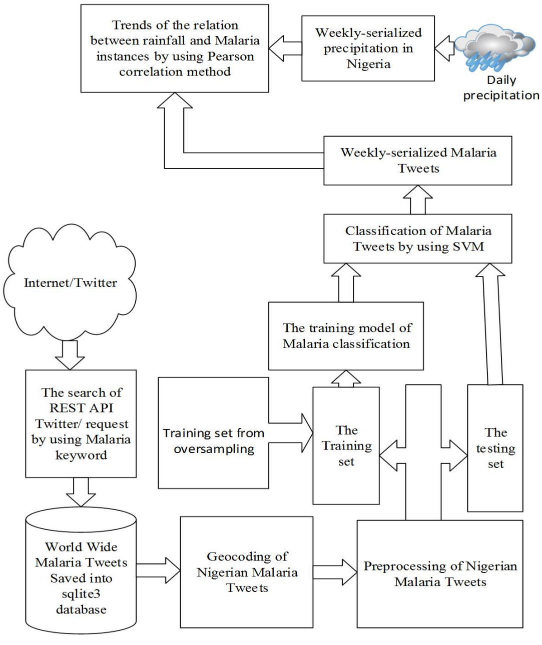

Trends of the relation between rainfall and Malaria instances by using Pearson correlation method
Weekly-serialized precipitation in Nigeria
Daily precipitation
Weekly-serialized Malaria Tweets
Classification of Malaria Tweets by using SVM
Internet/Twitter
The training model of Malaria classification
The search of REST API Twitter/ request by using Malaria keyword
Training set from oversampling
The Training set
The testing set
World Wide Malaria Tweets Saved into sqlite3 database
Geocoding of Nigerian Malaria Tweets
Preprocessing of Nigerian Malaria Tweets

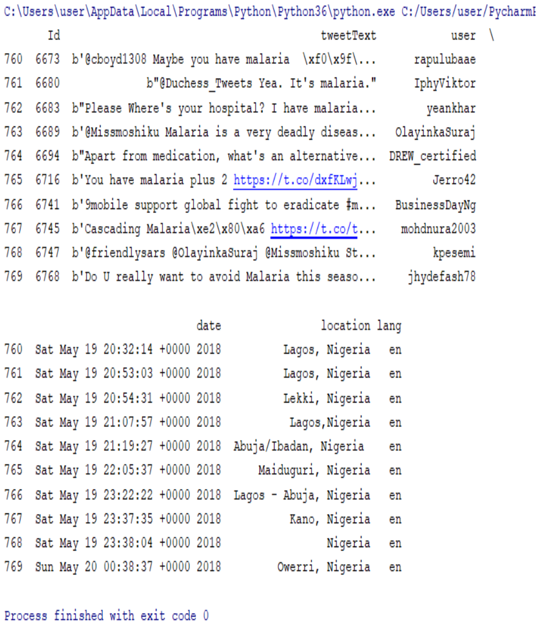
Fig. 1. Proposed model structure.

3.1 Twitter Streaming and Storage

We created an API account in order to have credential from Twitter and then we crawled worldwide tweets with 'malaria' filter keyword from May 8, till November 15, 2018. During that period of six months and one week we managed to crawl 19 Megabytes (Mbs) equivalent to 160000 rows of Malaria tweets worldwide. Those Tweets were stored as Java Script Object Notation (JSON) formatted data in Sqlite3 database.

3.2 Geocoding Nigerian Malaria Tweets

To select in the Nigerian Malaria Tweets from the entire worldwide Malaria dataset, we used a normal python program that took the worldwide Malaria dataset as input, and set 'nigeria' as selection condition along with other 51 locations (cities and states) in Nigeria which are likely to be in the location section of a typical Tweet, and filtered in Malaria Tweets in Nigeria only. This geocoding step gave us a dataset of 52000 rows of Malaria tweets in Nigeria. A sample of such Tweets is shown in Fig. 2.



```
C:\Users\user\AppData\Local\Programs\Python\Python36\python.exe C:/Users/user/Pycharm
       Id                                          tweetText           user  \
760  6673  b'@cboyd1308 Maybe you have malaria  \xf0\x9f\...     rapulubaae
761  6680               b"@Duchess_Tweets Yea. It's malaria."     IphyViktor
762  6683  b"Please Where's your hospital? I have malaria...       yeankhar
763  6689  b'@Missmoshiku Malaria is a very deadly diseas...  OlayinkaSuraj
764  6694  b"Apart from medication, what's an alternative...  DREW_certified
765  6716  b'You have malaria plus 2 https://t.co/dxfKLwj...        Jerro42
766  6741  b'9mobile support global fight to eradicate #m...  BusinessDayNg
767  6745  b'Cascading Malaria\xe2\x80\xa6 https://t.co/t...   mohdnura2003
768  6747  b'@friendlysars @OlayinkaSuraj @Missmoshiku St...        kpesemi
769  6768  b'Do U really want to avoid Malaria this seaso...     jhydefash78

                               date                location lang
760  Sat May 19 20:32:14 +0000 2018          Lagos, Nigeria   en
761  Sat May 19 20:53:03 +0000 2018          Lagos, Nigeria   en
762  Sat May 19 20:54:31 +0000 2018          Lekki, Nigeria   en
763  Sat May 19 21:07:57 +0000 2018           Lagos,Nigeria   en
764  Sat May 19 21:19:27 +0000 2018  Abuja/Ibadan, Nigeria    en
765  Sat May 19 22:05:37 +0000 2018      Maiduguri, Nigeria   en
766  Sat May 19 23:22:22 +0000 2018  Lagos - Abuja, Nigeria   en
767  Sat May 19 23:37:35 +0000 2018           Kano, Nigeria   en
768  Sat May 19 23:38:04 +0000 2018                 Nigeria   en
769  Sun May 20 00:38:37 +0000 2018         Owerri, Nigeria   en

Process finished with exit code 0
```

Fig. 2. Sample tweets after applying geocoding.

3.3 Twitter Data Preprocessing

We applied different Natural Language Processing Toolkits (NLTK) to clean out the undesirable characters from the Tweets. Those tools include regular expressions, stop words and duplicate removal, and stemming. After those preprocessing, we remained with 12425 Tweets ready to be classified into cases related to or unrelated to Malaria.

One of the important techniques in the preparation of Tweets for classification is stop words removal. Stop words in NLTK are English words often occurring repeatedly in sentence but carrying no meaningful information to the subject of a sentence. They are sub-sentences connectors, pronouns or some auxiliary verbs like 'do', 'have', 'can' and so on [15].

However, in some sentences classification, stop words can play an unavoidable role to infer the correct meaning of a sentence. Similarly, in our Tweets some stop words were kept in the sentences because of their role in the meaning of a sentence as a Malaria cases or not. For instance, the 'my Malaria is becoming worse' sentence indicates a Malaria cases because of the stop word 'my'. By contrast, if a Twitter user Tweets for example 'Malaria is becoming worse', without 'my', it becomes a general sentence which might mean Malaria in the whole region or something else. Such sentence would be classified unrelated to Malaria. Also 'I do have Malaria' and 'I do not have Malaria' are differently classified. Consequently 'do' and 'don't' are also not considered as stop words in our classification process. So, stop words like 'my' 'his', 'I', 'yours', 'have', 'do' and so on were kept in the Tweet sentence by creating a white list of stop words.

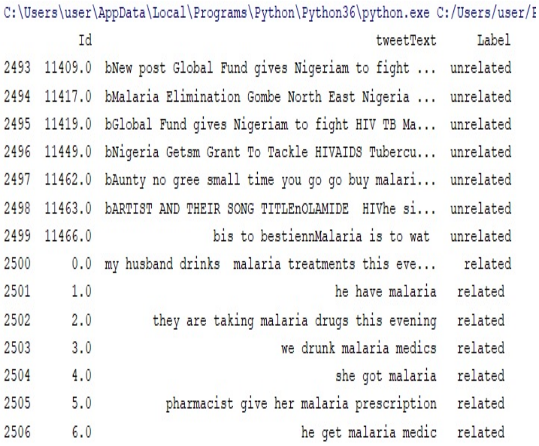
3.4 Oversampling Preprocessing

Tweets classification using supervised learning methods task is challenging in several respects. One of the most complicating aspect is the fact that tweets expressing an opinion about a given topic usually present a skewed polarity distribution. In this case, any classifier would be biased towards the majority class. In order to cope with this problem many projects which use Twitter as their source of data use oversampling to eliminate the imbalance into their data [16], [17]. One feature of a good classifier is having a relatively sufficient amount of training data equitably balanced according to the classes defined.

Our training data set was also unbalanced. Tweets manually labeled as Malaria case was scarce compared to other Twitter messages just mentioning Malaria disease without necessarily telling a Malaria patient. We solved it by using a random oversampling technique. To apply it, we first of all studied the main vocabularies in the sentences labeled Malaria cases. Then we built a python program that combines those vocabularies in, names, pronouns and verbs into a meaningful random sentence. With that program we were able to generate any amount of different sentences related to Malaria. As we were aware of the potential flaw of repetition of the same random sentence with this technique, we also used a duplicate sentence removal module and eliminated multiplicatively generated random sentences.

3.5 Training Model and Classification

A Tweet was labeled related to Malaria instance if it talks about someone suffering from Malaria, either Twitter user himself or the user talking about someone else having the Malaria but in his direct vicinity: for example if they are in the same family or place. For instance 'My girlfriend is on Malaria drugs' or 'We have a Malaria patient in our village'. The Tweet also has to be reported not from a far past time. We accepted a Tweet within one week maximum. For instance 'She recovered from Malaria' would be labeled unrelated to Malaria patient but 'I stopped Malaria drugs a week ago' would be a Malaria instance Tweet. We manually labeled 205 Malaria related Tweets, generated randomly 2295 Malaria related Tweets and combined with 2500 manually labeled Malaria unrelated and we used a 5000 Tweets of training against 12425 unseen Tweets to be classified. A sample of those training Tweets and their two labels are shown in Fig. 3.



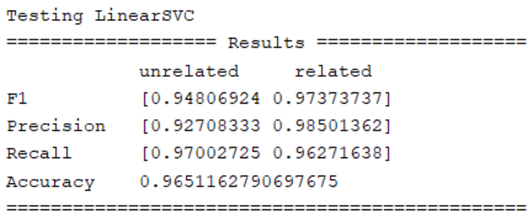
```
C:\Users\user\AppData\Local\Programs\Python\Python36\python.exe C:/Users/user/E
           Id                                          tweetText      Label
2493  11409.0  bNew post Global Fund gives Nigeriam to fight ...  unrelated
2494  11417.0  bMalaria Elimination Gombe North East Nigeria ...  unrelated
2495  11419.0  bGlobal Fund gives Nigeriam to fight HIV TB Ma...  unrelated
2496  11449.0  bNigeria Getsm Grant To Tackle HIVAIDS Tubercu...  unrelated
2497  11462.0  bAunty no gree small time you go go buy malari...  unrelated
2498  11463.0  bARTIST AND THEIR SONG TITLEnOLAMIDE  HIVhe si...  unrelated
2499  11466.0                   bis to bestiennMalaria is to wat  unrelated
2500      0.0  my husband drinks  malaria treatments this eve...    related
2501      1.0                                   he have malaria   related
2502      2.0             they are taking malaria drugs this evening   related
2503      3.0                            we drunk malaria medics   related
2504      4.0                                   she got malaria   related
2505      5.0           pharmacist give her malaria prescription   related
2506      6.0                               he get malaria medic   related
```

Fig. 3. Sample tweets used in training.

Then we built four different classification models including Bernoulli Naive Bayes test classifier denoted as NB, Random Forest (RF), Support Vector Machine (SVM) and Xgboost (XGB), all with a 7:3 training-test ratio. We also tried 8-folds cross-validation for SVM, NB and RF. Support Vector Machine performed well and scored an accuracy of 96% as can be seen in Fig. 4. The results for different classification models are shown in Table 1. From that table we can visualize that ac-

Table 1. Comparison of performance of classification modules in %

Classifier	F1-Score	Recall	Precision	Accuracy
NB	95	91	99	94
NBcv	-	-	-	92
RF	93	88	98	91
RFcv	-	-	-	0.88
XGB	91	95	87	87
SVM	97	96	98	96
SVNcv	-	-	-	94



```
Testing LinearSVC
=================== Results ===================
               unrelated       related
F1             [0.94806924 0.97373737]
Precision      [0.92708333 0.98501362]
Recall         [0.97002725 0.96271638]
Accuracy       0.9651162790697675
==============================================
```

Fig. 4. SVM classification results.

curacies of all tried classifiers are promising. The table also records the precision, recall and F1-score performance metrics for a Tweet to be classified as Malaria case as shown in equations (2), (3) and (4) respectively.

The performance of a machine learning classifier can be evaluated by computing the following metrics. The number of correctly recognized class examples known as True Positives, the number of correctly recognized examples that do not belong to the class known as True Negatives, and examples that either were incorrectly assigned to the class known as False Positives or that were not recognized as class examples known as False Negatives. These four counts constitute a confusion matrix as shown in Table 2 for the case of binary classification [17] as we used it in our paper. Accuracy, precision, recall and F1- Score are based on the data in the confusion matrix.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}, \tag{1}$$

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

$$F1-Score = \frac{2*Recall*Precision}{Recall+Precision} \tag{4}$$

Supervised machine learning performances are typically a function of their accuracy. One of the key objectives in classification is often to achieve a high accuracy. Accuracy approximates how effective the algorithm is by showing the probability of the true value of the class label as shown in equation (1). It assesses the overall effectiveness of the algorithm and it sometimes suffices enough the classification evaluation especially when the classification is binary as in our case [19], [19]. Precision tries to answer what proportion of positive identification was actually correct as indicated in (2). Recall defines the proportion of actual positive was identified correctly as in (3) while the F1-Score combines precision and recall relative to a specific class as in (4). F1-Score can also be interpreted as a weighted average of the precision and recall.

Table 2. Confusion Matrix for binary classification

Data class	Classified positive	Classified negative
Positive	True positive(TP)	False Negative(FN)
Negative	False Positive (FP)	True Negative(TN)

The SVM classifier gave us a total of 596 Malaria instances in the whole Nigeria from May 8 till 15 November 2018. Those Malaria instances were counted day by day and then we serialized them on a week and monthly basis. The monthly trend of those Malaria instances are shown in Fig. 5. As it could be visualized the Malaria patients rose gradually from May, reached pick in September and then started decreasing.

3.6 Precipitation Volume Data

Nigeria has a tropical climate with two seasons;

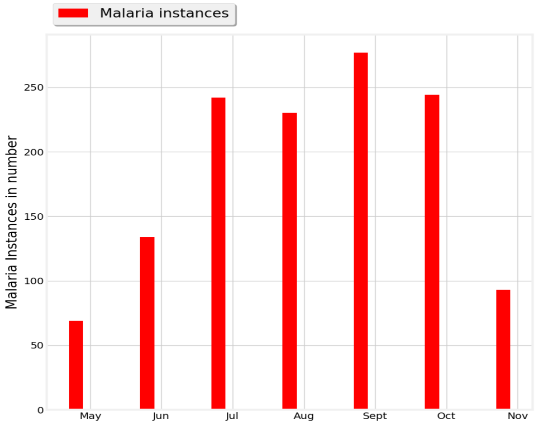
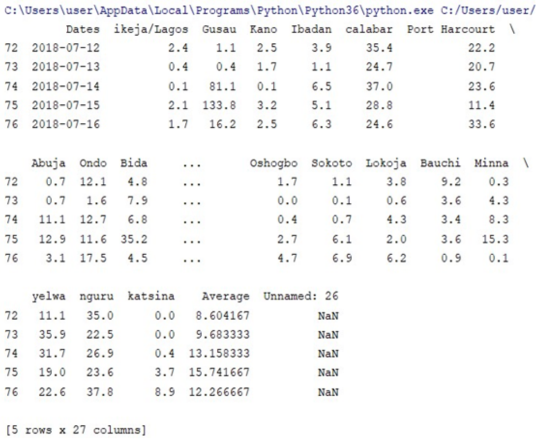


Fig. 5. Monthly trend of malaria cases.

wet and dry. The wet or rainy season commences from May and terminates early November. It highly affects the development of Anopheles which propagates Malaria [13]. The precipitation volume we compared with Malaria instances both in Nigeria is an average precipitation calculated over daily rain volume in 24 important climatic regions in Nigeria. We defined the main rainfall regions by referring to the work done by Akinsanola A.A and K. Ogunjobi on the analysis of rainfall and temperature variability over Nigeria [21]. From those samples we observed that daily precipitations in July are relatively the highest in many regions. A reader may refer to the Fig. 6 and Fig. 7 for the samples of daily precipitation in July and November respectively. The average was computed just by using an excel function 'average' and it was computed over the 24 regions. We then summed and serialized the daily average precipitation over week and month. The monthly trend of precipitation are shown in Fig. 8. As explained in the introduction, the trend shows an increase in precipitation from May till July and then started decreasing from August and it is comparable to the typical rainfall pattern in the whole Nigeria.



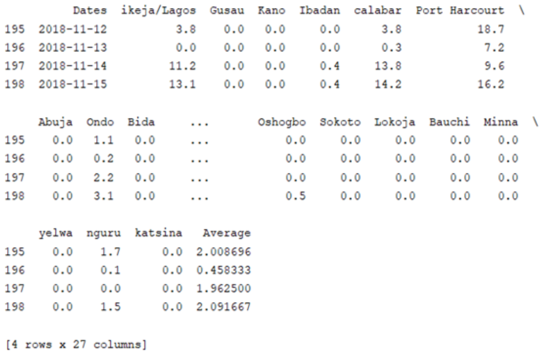
```
C:\Users\user\AppData\Local\Programs\Python\Python36\python.exe C:/Users/user/
         Dates  ikeja/Lagos  Gusau  Kano  Ibadan  calabar  Port Harcourt  \
72  2018-07-12          2.4    1.1   2.5     3.9     35.4           22.2
73  2018-07-13          0.4    0.4   1.7     1.1     24.7           20.7
74  2018-07-14          0.1   81.1   0.1     6.5     37.0           23.6
75  2018-07-15          2.1  133.8   3.2     5.1     28.8           11.4
76  2018-07-16          1.7   16.2   2.5     6.3     24.6           33.6

    Abuja  Ondo  Bida     ...     Oshogbo  Sokoto  Lokoja  Bauchi  Minna  \
72    0.7  12.1   4.8     ...         1.7     1.1     3.8     9.2    0.3
73    0.7   1.6   7.9     ...         0.0     0.1     0.6     3.6    4.3
74   11.1  12.7   6.8     ...         0.4     0.7     4.3     3.4    8.3
75   12.9  11.6  35.2     ...         2.7     6.1     2.0     3.6   15.3
76    3.1  17.5   4.5     ...         4.7     6.9     6.2     0.9    0.1

    yelwa  nguru  katsina    Average  Unnamed: 26
72   11.1   35.0      0.0   8.604167          NaN
73   35.9   22.5      0.0   9.683333          NaN
74   31.7   26.9      0.4  13.158333          NaN
75   19.0   23.6      3.7  15.741667          NaN
76   22.6   37.8      8.9  12.266667          NaN

[5 rows x 27 columns]
```

Fig. 6. Sample daily precipitation in 24 main regions in Nigeria recorded in a typical July.



```
          Dates  ikeja/Lagos  Gusau  Kano  Ibadan  calabar  Port Harcourt  \
195  2018-11-12          3.8    0.0   0.0     0.0      3.8           18.7
196  2018-11-13          0.0    0.0   0.0     0.0      0.3            7.2
197  2018-11-14         11.2    0.0   0.0     0.4     13.8            9.6
198  2018-11-15         13.1    0.0   0.0     0.4     14.2           16.2

     Abuja  Ondo  Bida     ...     Oshogbo  Sokoto  Lokoja  Bauchi  Minna  \
195    0.0   1.1   0.0     ...         0.0     0.0     0.0     0.0    0.0
196    0.0   0.2   0.0     ...         0.0     0.0     0.0     0.0    0.0
197    0.0   2.2   0.0     ...         0.0     0.0     0.0     0.0    0.0
198    0.0   3.1   0.0     ...         0.5     0.0     0.0     0.0    0.0

     yelwa  nguru  katsina   Average
195    0.0    1.7      0.0  2.008696
196    0.0    0.1      0.0  0.458333
197    0.0    0.0      0.0  1.962500
198    0.0    1.5      0.0  2.091667

[4 rows x 27 columns]
```

Fig. 7. Sample daily precipitation in 24 main regions in Nigeria recorded in a typical November.

3.7 Calculation of Relation between Serial Malaria Instances and Precipitations in Nigeria

Correlation between two variables represents the strength of the putative linear association between the variables in question. It is a dimensionless quantity that takes a value in the range -1 to 1. A correlation coefficient of zero indicates that no linear relationship exists between two continuous variables. But a correlation of -1 or 1 indicates that the two variables are associated but not causal. If the coefficient of correlation is a positive number it means that the two variables increases or decreases similarly following same pattern. By contrast the correlation is negative if one variable increases while the other is decreasing or vice versa. Pearson and Spearman correlation coefficients are the two main statistical measures of

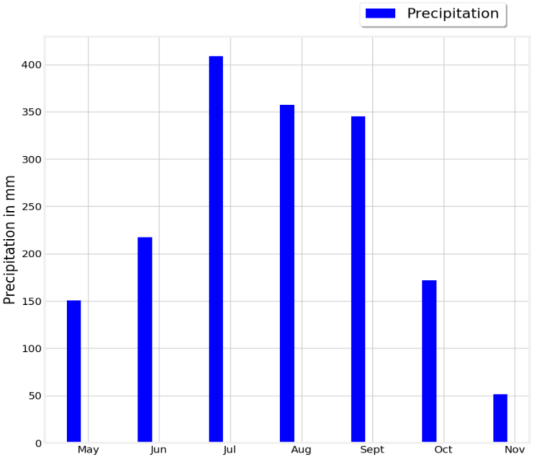


Fig. 8. Monthly trend of precipitation volume.

correlations. Pearson correlation is used when the two variables are of normally distributed data and when none of them is skewed. It also requires that none of the data to be compared are extreme data because it tends to be biased in that case. Our serial data meet all those requirements for the Pearson correlation coefficient. The Spearman correlation coefficient is used mainly when the data are skewed and it is a rank coefficient [22].

We calculated the correlation between the serial data of Malaria and those of precipitation by using a python program. The “corr (method=Pearson)” function we used is found in searbon module of python. As we explained in the introduction of this paper, the effect of precipitation on the Anopheles development appear after 10 to 20 days or just a two weeks period interval. Keeping that in mind we compared Malaria instances with precipitations starting before as 2 weeks, one week and also starting at the same date. That is, for a two weeks-before comparison, we matched the sum of daily Malaria instances from May 15 till May 22 with the sum of averages of daily precipitation taken from May 1 till May 8 and so on. For the one week-before comparison, we matched the sum of daily Malaria instances from May 15 till May 22 with the sum of average precipitations taken from May 9 till May 15 and so on. And for the same-week comparison we matched the sum of daily Malaria instances from May 8 till May 15 with the sum of average precipitations taken from May 8 till May 15 and so on. The relational results in two weeks are shown in Fig. 9, the relation result in one week difference is in Fig. 10 and the relation when the Malaria instances and precipitations are of the same date are shown in Fig. 11. We also showed the trend in the relation of those variables by month and it can be seen in Fig. 12. Along with those combined trends we also calculated the correlation coefficient matrix to show how the two variables are correlated by coefficient value.

4. RESULTS AND DISCUSSION

In our research we fetched Malaria instances in Nigeria by crawling, preprocessing and classifying Malaria related messages from Twitter. Then we compared those Malaria instances with precipitation average calculated over 24 climatically main regions in Nigeria recorded daily from an internet Website as could be referred to in [23].

We calculated the Pearson coefficient of correlation between weekly Malaria serial data and precipitation serial data in two weeks before, one week

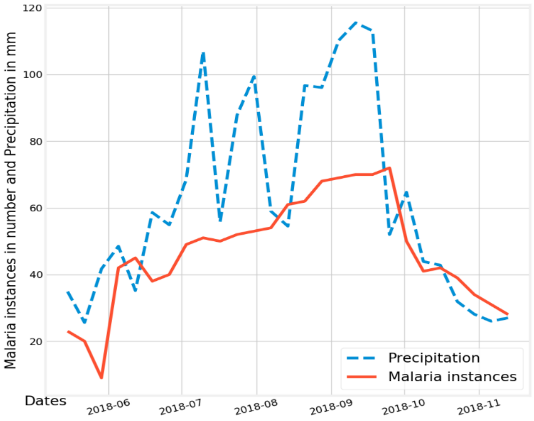


Fig. 9. Trend in relation between malaria instances from twitter and precipitation recorded two weeks before in Nigeria.

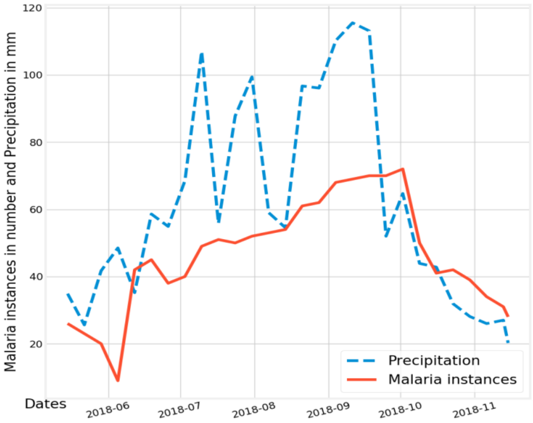


Fig. 10. Trend of relation between malaria instances from twitter and precipitation recorded one week before in Nigeria 2018.

before and same week. The results of correlations showed that correlation between Malaria and precipitation recorded two-weeks before occurrence of Malaria is 0.75 and is the highest. Also from Fig. 9 we can see that the trend of variation of Malaria and precipitation have fairly a similar pattern even though there are some minor differences in variations. The correlation when the precipitation are in 1 week before is 0.69 and the variation trend in Fig. 10 also shows a similar variation pattern. The comparison when both Malaria and precipitation vary starting from the same date gave a correlation coefficient of 0.55. The monthly trend of Malaria and precipitation is also shown in Fig. 12. As described in the introduction on the pattern of the rain fall in Nigeria which is recorded high from June and decreasing from October, we also noticed a same pattern. Our Malaria instances from Twitter vary following a similar trend with an exception in September where the Malaria fell while precipitation was rising.

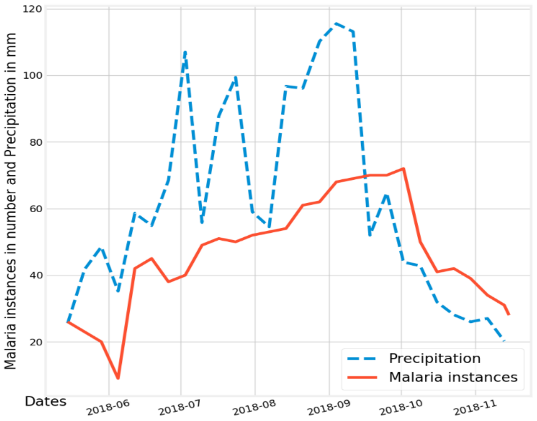


Fig. 11. Trend in relation between malaria instances from twitter and precipitation recorded at same weeks in Nigeria 2018.

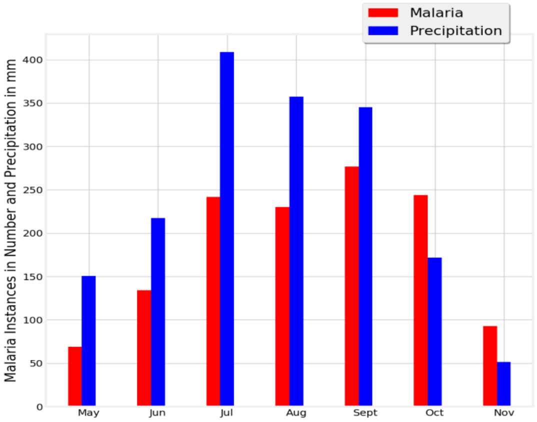


Fig. 12. Trend in relation between malaria instances from twitter and precipitation by months 2018.

5. CONCLUSION

Our research proposed a method for monitoring and reporting Malaria instances by using Malaria Twitter data and precipitation. We live-streamed Nigerian Twitter messages mentioning Malaria. Then we classified those messages into two classes namely Malaria case related and Malaria case unrelated classes by using SVM. The results of the class related to Malaria was then compared with average precipitation during the same period in Nigeria. The comparison results show a high correlation coefficient between those two variable. Therefore we can state that Twitter data can be used to directly monitor and report on Malaria instances. We also can conclude that Malaria outbreak can be predicted basing on the precipitation conditions especially in Nigeria and also some other parts of the SSA where Malaria is endemic.

We were not able to conduct our project during a whole year of 12 months or more because of limited time of our research schedule. As it would further improve correlation coefficient results, we plan to do it in our future research. Finally, we call for the health care especially those working in Malaria area to use SNSs to monitor and report on Malaria outbreaks and to include precipitation volume and seasonality in their planning about this disease so that we can eliminate it completely.

REFERENCES

[1] World Health Organization, World Malaria Report, 2018.

[2] C.W. Gitonga, P.N. Karanja, J. Kihara, M. Mwanje, E. Juma, R.W. Snow, et al., "Implementing School Malaria Surveys in Kenya: Towards a National Surveillance System," *Malaria Journal*, Vol. 9, No. 1, pp. 306, 2010.

[3] M.I. Joo, D.H. Ko, and H.C. Kim, "Development of Smart Healthcare Wear System for Acquiring Vital Signs and Monitoring Personal Health," *Journal of Korea Multimedia Society*, Vol. 19, No. 5, pp. 808- 817, 2016.

[4] J. Boit and H. Alyami, "Malaria Surveillance System using Social Media," *Proceedings of the 13th Midwest Association for Information Systems (MWAIS) Conference*, 2018.

[5] R. Xuriguera, *Using Twitter as a Source of Information for Time Series Prediction*, Master's Thesis of Catalunya Polytechnic University, 2012.

[6] S. Michal and A. Romanowski, "Sentiment Analysis of Twitter Data Within Big Data Distributed Environment for Stock Prediction," *Proceeding of 2015 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pp. 1349-1354, 2015.

[7] M.Y. Anwar, J.A. Lewnard, S. Parikh, and V.E. Pitzer, "Time Series Analysis of Malaria in Afghanistan: Using ARIMA Models to Predict Future Trends in Incidence," *Malaria Journal*, Vol. 15, No. 1, pp. 566, 2016.

[8] S. Hundessa, G. Williams, S. Li, J. Guo, W. Zhang, and Y. Guo, "The Weekly Associations Between Climatic Factors and Plasmodium Vivax and Plasmodium Falciparum Malaria in China 2005-2014," *Transactions of the Royal Society of Tropical Medicine and Hygiene*, Vol. 111, No. 5, pp. 211-220, 2017.

[9] G.J. Abiodun, R. Maharaj, P. Witbooi, and K.O. Okosun, "Modelling the Influence of Temperature and Rainfall on the Population Dynamics of Anopheles Arabiensis," *Malaria Journal*, Vol. 15, No. 1, pp. 364, 2016.

[10] R. Ranovan, A. Doewes, and R. Saptono, "Twitter Data Classification Using Multinomial Naive Bayes for Tropical Diseases Mapping in Indonesia," *Journal of Telecommunication, Electronic and Computer Engineering*, Vol. 10, No. 2-4, pp. 155-159, 2018.

[11] H. Achrekar, A. Gandhe, R. Lazarus, S.H. Yu, and B. Liu, "Predicting Flu Trends using Twitter Data," *Proceedings of 2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pp. 702-707, 2011.

[12] A. Ayanlade, N.O. Adeoye, O. Babatimehin "Intra-Annual Climate Variability and Malaria Transmission in Nigeria," *Bulletin of Geography. Socio - economic Series*, Vol. 21, No. 21, pp. 7-19, 2013.

[13] K. Olayemi, A.T. Ande, A.V. Ayanwale, A.Z. Mohammed, T.M. Bello, B. Idris, et al., "Seasonal Trends in Epidemiological and Entomological Profiles of Malaria Transmission in North Central Nigeria," *Pakistan Journal of Biological Sciences*, Vol. 14, No. 4, pp. 293-299, 2011.

[14] A.B. Adigun, E.N. Gajere, O. Oresanya, and P. Vounatsou, "Malaria Risk in Nigeria: Bayesian Geostatistical Modelling of 2010 Malaria Indicator Survey Data," *Malaria Journal*, Vol. 14, No. 1, pp. 156, 2015.

[15] A. Bakliwal, P. Arora, S. Madhappan, N. Kapre, M. Singh, and V. Varma, "Mining Sentiments from Tweets," *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, pp. 11-18, 2012.

[16] J. Ah-Pine and E.P.S. Morales, "A study of Synthetic Oversampling for Twitter Imbalanced Sentiment Analysis," *Processing of Workshop on Interactions Between Data Mining and Natural Language*, pp. 17-24, 2016.

[17] X. Zhang, X. Lin, J. Zhao, Q. Huang, and X. Xu, "Efficiently Predicting Hot Spots in PPIs

by Combining Random Forest and Synthetic Minority Over-Sampling Technique," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pp. 1-1, 2018.

[18] M. Sokolova and G. Lapalme, "A Systematic Analysis of Performance Measures for Classification Tasks," *Information Processing and Management*, Vol. 45, No. 4, pp. 427-437, 2009.

[19] I. Ahmad, M. Basheri, M.J. Iqbal, and A. Raheemm, "Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for Intrusion Detection," *IEEE Access*, Vol. 6, pp. 33789-33795, 2018.

[20] A. Mathur and G.M. Foody, "Multiclass and Binary SVM Classification: Implications for Training and Classification Users," *IEEE Geoscience and Remote Sensing Letters*, Vol. 5, No. 2, pp. 241-245, 2008.

[21] K. Akinsanola A.A, and Ogunjobi K.O, "Analysis of Rainfall and Temperature Variability Over Nigeria," *Global Journal of Human-Social Science Research*, Vol. 14, No. 3, pp. 10-28, 2014.

[22] M.M. Mukaka, "A Guide to Appropriate Use of Correlation Coefficient in Medical Research," *Malawi Medical Journal*, Vol. 24, No. 3, pp. 69-71, 2012.

[23] World Weather Online, https://www. worldweatheronline.com/weather-history/federal-capital-territory/ng.aspx (Accessed Oct. 1-Nov 15, 2018).

Nduwayezu Maurice

2012 Information and Communication Systems Engineering, National University of Rwanda, Rwanda (B.S)
2019 Information and Communications Systems Engineering, Inje University, Korea (M.S)
2019~Current Department of Information and Communication Systems, Inje University, M.S

Satyabrata Aich

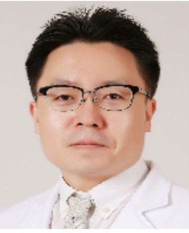
2008 Mechanical Engineering IIT Madras, India (M.Tech)
2019 Computer Engineering, Inje University (Ph.D)
2019~currently working as a Research Professor, Institute of Digital Anti-aging Healthcare (IDA), Inje University

Han Suk Young

1999 Busan National University Computer Engineering Department (Bachelor)
2003 USC, Software Engineering (Master)
2004~2014 Product Planning for Mobile Phone in Samsung Electronics" Wireless Business Department
2012~2017 The Way Consulting CEO
2017~Currently Professor of Institute or Digital Anti-aging Healthcare at Inje University

Kim Jung Eon

2006 Gyeongsang National University, School of Medicine (M.D.)
2018~Current Inje University Ilsan Paik Hospital, Department of Emergency Medicine, Clinical Professor

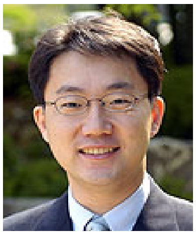


Kim Hoon

2002 Inha University, College of Medicine (M.D.)
2007 Inha University, College of Medicine, Graduate school, (M.S.)
2017 Inha University, College of Medicine, Graduate school (Ph.D.)
2010~Current Inje University Ilsan Paik Hospital, Department of Emergency Medicine, Professor

Park Junseok

1997 Yonsei Wonju Medical School (M.D.)
2008 Yonsei University College of Medicine, Graduate School (M.S.)
2014 Chungnam Univeristy College of Medicine, Graduate School (Ph.D.)
2005~Current Inje University Ilsan Paik Hospital, Department of Emergency Medicine, Professor



Hwang Won-Joo

1998 Pusan National University Dept. of Computer Engineering (B.E.)
2000 Pusan National University Dept. of Computer Engineering (M.E.)
2002 (Japan) Osaka University Dept. of Information Systems Engineering (Ph.D.)
2002~Current Inje University Dept. Of Electronic, Telecommunications, Mechanical & Automotive Engineering, Professor