Hindawi
Complexity
Volume 2021, Article ID 5520366, 12 pages
https://doi.org/10.1155/2021/5520366

Research Article

Early Detection of Seasonal Outbreaks from Twitter Data Using Machine Learning Approaches

Samina Amin,[1] Muhammad Irfan Uddin,[1] Duaa H. alSaeed,[2] Atif Khan ,[3] and Muhammad Adnan[1]

[1]*Institute of Computing, Kohat University of Science and Technology, Kohat 26000, Pakistan*
[2]*College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia*
[3]*Department of Computer Science, Islamia College Peshawar, Peshawar 25000, Pakistan*

Correspondence should be addressed to Atif Khan; atif.softeng@gmail.com

Received 1 February 2021; Revised 22 February 2021; Accepted 26 February 2021; Published 15 March 2021

Academic Editor: Furqan Aziz

Seasonal outbreaks have several different periods that occur primarily during winter in temperate regions, while influenza may occur throughout the year in tropical regions, triggering outbreaks more irregularly. Similarly, dengue occurs in the star of the rainy season in early May and reaches its peak in late June. Dengue and flu brought an impact on various countries in the years 2017–2019 and streaming Twitter data reveals the status of dengue and flu outbreaks in the most affected regions. This research work presents that Social Media Analysis (SMA) can be used as a detector of the epidemic outbreak and to understand the sentiment of social media users regarding various diseases. Providing awareness about seasonal outbreaks through SMA is an effective approach for researchers and healthcare responders to detect the early outbreaks. The proposed model aims to find the sentiment about the disease in tweets, and the seasonal outbreaks-related tweets are classified into two classes as disease positive and disease negative. This work proposes a machine-learning-based approach to detect dengue and flu outbreaks in social media platform Twitter, using four machine learning algorithms: Random Forest (RF), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Decision Tree (DT), with the help of Term Frequency and Inverse Document Frequency (TF-IDF). For experimental analysis, two datasets (dengue and flu) are analyzed individually. The experimental results show that the RF classifier has outperformed the comparison models in terms of improved accuracy, precision, recall, F1-measure, and Receiver Operating Characteristic (ROC) curve. The proposed work offers favorable performance with total precision, accuracy, recall, and F1-measure ranging from 84% to 88% for conventional machine learning techniques.

1. Introduction

An infectious disease may have several periods that typically occur in a specific season in the prevaccination era. Rapid identification of a seasonal outbreak is essential to generate a reaction to healthcare professionals more quickly and efficiently. The seasonal outbreaks can lead to serious diseases, such as influenza or Influenza-Like Illness (ILI), which can lead to death when the epidemic breaks down epidemiologically in a region [1, 2]. Influenza is a respiratory system disease, which causes a significant death rate every year globally. The flu virus is often medically mild and is acknowledged by symptoms such as headache, sneeze, fever, sore throat, and cough [3]. Influenza shots are almost always available during the winter season, and the infected person should move to a specialist instead of a normal doctor. The barrier of influenza can harm the patient and can create a much severe condition if not treated. As reported in [4], flu is an epidemic outbreak, and in such cases where the proliferation of infectious diseases, particularly influenza, is a real risk to people, the government must implement appropriate health surveillance to control the epidemic. Similarly, dengue infection is a mosquito-borne virus that causes serious ILI and also causes a potentially fatal risk factor called severe infection with dengue fever. Dengue is among the rapidly propagating contagious diseases in the world. Therefore, providing real-time monitoring, early detection, and identification of contagious diseases related to the outbreak of influenza or dengue are important for public health [1, 5].

In the health surveillance systems, Online Social Media (OSM) offers effective resources for epidemic outbreaks detection and an active way of coping against the outbreaks [6, 7]. The effect of seasonal epidemic outbreaks (i.e., dengue or influenza) on population safety can be minimized by early notice of disease detection. To track the frequency of infectious outbreaks faster than healthcare practitioners and health agencies such as the Center for Disease Control and Prevention (CDC), OSM can also be configured for surveillance systems [1]. The CDC uses the Influenza-Like-Illness Surveillance Network (ILINet), a platform used to track early alerts of flu outbreaks by healthcare professionals. While it is an effective yet expensive and slow process, it requires weeks or even months when data becomes accessible from CDC. Therefore, various studies concentrate on offering alternatives to track ILI through leveraging SMA and identifying early alerts regarding infectious diseases to conduct real-time analysis. OSM applications like Twitter can be used to predict public disease outbreaks and can promote timely information [8–10]. With the help of OSM, it is possible to alert healthcare consultants to provide the appropriate services and monitor the epidemic. Nowadays, people frequently use OSM applications to share ideas, opinions, and health status, specifically when there is an epidemic in a region. In order to decrease epidemic outbreaks, SMA can be used to deliver efficient information for disease monitoring and is a useful way to interact with the public [11–17].

In this paper, a machine-learning-based intelligent model is proposed, which will retrieve text on dengue and flu from Twitter. The tweets are categorized into two groups, that is, positive and negative, where positive tweets represent dengue and flu-infected cases (such as to represent the symptoms of dengue and flu in infected people in tweets) and negative tweets represent non-dengue nor flu-infected cases. The main contributions of the proposed work are the use of machine learning techniques RF, KNN, SVM, and DT for tweet identification with the help of a TF-IDF approach. Applying a unigram approach, which means deriving sentiment of the public on dengue from an individual word in tweets, the early detection of seasonal outbreaks (dengue and flu in our case) through SMA can provide awareness about these seasonal outbreaks. It identifies dengue- and flu-related tweets in various regions using tweet data. Some people still think dengue is not dangerous and consider it similar to a "seasonal flu" which is a seasonal disease with available treatments and vaccines, not to mention that even seasonal flu kills around 30 k Americans a year.

The rest of the paper is structured as follows: Section 2 presents a brief review of related work. The approach followed to conduct the experimental results is confirmed in Section 3, while in Section 4, the evaluation and the analysis of the results are presented. The concluding remarks and future research lines are presented in Section 5.

2. Related Work

In this section, a review of related work conducted on detecting seasonal epidemics from social media is presented.

With the enhancement of machine learning, it has been increasingly adopted for the SMA regarding disease detection. Wang et al. [18] developed a framework for influenza prediction based on real-time online OSM data. In their work, they deployed a Partial Differential Equation (PDE) for prediction. Furthermore, with flu reduction evaluation, they further predicted the volume of the tweets in the future. Chen et al. [19] presented a model based on two temporal topic models such as supervised and unsupervised models to grab the user's hidden states and geographic information from their tweets message for the purpose of better trends estimation. The gap between surveillance strategies for phenomenological disease and epidemiological techniques has been narrowed by this approach using tweet data.

Recently, the influenza outbreak has been mainly triggered by CDC, which is now spreading all over the year, triggering a rise in instances between January and March. CDC has been encouraged by professionals to perform its role in supporting awareness regarding early detection and provide necessary recourses to control or cure flu epidemic that has become a risk for people with the advent of cold weather. There was a need for media campaigns for awareness on the flu shot, and the healthcare organizations concerned must initiate efforts to educate people regarding the impacts of flu on their life preventative action they can take and therefore the treatments and appropriate medicines [20]. Now, in the OSM era, disease-related information is also shared on OSM directly or indirectly [2]. To utilize the information about flu outbreak, it can be easily extracted from SMA, particularly microblogging platform Twitter, to detect the early outbreak of influenza shot for the awareness of healthcare professionals to provide resources or medications and to control the epidemic [1, 2].

The dengue outbreak is a globally transmitting contagious infection [21]. In order to effectively detect the outbreak of dengue fever and to examine the effect on primary prevention, dengue monitoring data is highly needed [22]. In China, dengue tends to be a significant public health concern with enlarging regions and massive cases recently [23]. However, in the scenario of infectious diseases and dengue virus surveillance systems in the big country of China, no extensive steps have been taken to predict and monitor early warnings of the dengue outbreak so far [23].

With powerful cooperative reaction from government and Nongovernmental Organizations (NGOs), healthcare professionals are helping to prevent further spread of the epidemic [24]. Public awareness about the seasonal epidemic outbreaks, especially regarding dengue fever, actual problem understanding, and the approaches to monitor dengue outbreaks are significant considerations. The perceptions, behaviour, and approaches regarding dengue in cities are explored in many studies [4, 25]. Social networking could be utilized efficiently to classify people contaminated with diseases and health awareness influences (e.g., influenza, dengue fever, anxiety, malaria, measles, etc.) with an intervention to improve public health. Through the use of SMA, emergency alert signs of the seasonal outbreaks can be identified and the time between occurrence and diagnosis can thus be shortened. Table 1 demonstrates the relevant research studies on outbreak detection using machine learning approaches.

The literature has shown that the previous studies on OSM-based disease detection focused on conventional machine learning approaches including Naïve Bayes, Support Vector Machine, and linear regression; and many works focus on detecting the frequency of tweets about a disease. The important considerations of the proposed work are the primary prevention and intervention that provide identification and alert system of infectious disease outbreaks, epidemic tracking, and modelling and evaluation of public health emergency. Recent years have observed a fast development of machine learning approaches in rapidly changing, dynamic, and data-rich environments to achieve these tasks. In this article, a set of current publications based on how to improve the use of the virtual world's sensor and OSM data to enhance seasonal outbreaks detection and early warning capacities are summarized. Furthermore, a collection of methods aimed at mapping the spread of seasonal outbreaks and estimating epidemiological data from Twitter messages are also presented. The proposed work focuses on supervised machine learning approaches DT, KNN, SVM, and RF and how to improve the performance of the traditional machine learning approaches to analyze seasonal epidemic outbreaks. The proposed work presents promising results for traditional machine learning techniques.

3. The Proposed Methodology

The proposed methodology incorporates various components including data gathering, preprocessing, feature extraction, and classifier. Figure 1 represents the proposed methodology adopted for dengue detection. In the following subsections, we will discuss each component in detail.

3.1. Data Gathering. In this work, the benchmark dataset on dengue and flu designed by Amin et al. [10] is analyzed. They labelled 6000 tweets on dengue and flu as infected or not infected. Tweets are considered as positive when they represent the symptoms of dengue\flu in infected people and the remaining tweets are considered as negative when they represent dengue-\flu-related information but not symptoms as shown in Table 2, while Table 3 shows the number of total labelled tweets.

3.2. Preprocessing. After a text is obtained, the collected data is promoted to certain preprocessing steps usually applied in natural language processing techniques [30] such as eliminating and removing stop words, sparse terms, and particular words, as those do not convey meaningful information. Then punctuations, accent marks, and other diacritics were removed. After that, the tweet text was converted into token words. These preprocessing steps were incorporated in order to enhance the efficiency of the proposed model and to improve the processing speed.

3.3. Feature Extraction. To convert text data into numbers, the machine learning techniques feature vectors and TF-IDF are adopted in this work. Feature vector converts a tweet into a matrix of token counts, while TF-IDF is the most common technique for feature selection in machine learning [31]. In a given corpus or dataset, the TF-IDF calculation indicates the significance of a term. By weighting the frequency of occurrence in the text and calculating how much the same word appears in other texts, TF-IDF calculates how significant a word is. If a word appears in a specific document several times but not in others, then it may be extremely important to that particular document and thus more significance is given. Mathematically, it is calculated as follows:

$$\begin{aligned}\mathrm{TF} &= \frac{(\text{number of occurrences of a word in a document})}{(\text{total of words in a document})},\\ \mathrm{IDF} &= \log\left[\frac{(\text{total number of documents})}{(\text{number of documents containing the word})}\right].\end{aligned} \tag{1}$$

The definition of the equations stated above is as follows: Term Frequency (TF): it calculates the number of occurrences of a word in certain tweets. Inverse Document Frequency (IDF): the calculation of TF-IDF indicates the significance of a word in a given document.

4. Machine Learning Models

Four machine learning algorithms (DT, KNN, SVM, and RF) are applied and evaluated for the classification task. In the following subsections, we will discuss those algorithms and explain the evaluation measures used.

5. Decision Tree

One of the foremost machine learning techniques is the Decision Tree (DT) [32]. DT can be used to solve both classification and regression problems, but it is most commonly used to tackle classification problems. It builds classification models in the form of a tree-like structure, where internal nodes represent the attributes of a dataset, root represents the decision rules, and each leaf node represents the outcome. The DT tree nodes consider the various levels, in which the root node is considered the first or top-most node. All inner networks contain measures on input variables or attributes (i.e., nodes having at least one child). The classification model splits towards the relevant child node, based on the test result, where the test and splitting process continues until it hits the leaf node. The nodes of the leaf or terminal match the results of the decision. DTs have been observed to be simple to understand and easy to learn and are a basic feature of many procedures for medical diagnosis. The results of all the tests for each node across the path can

Table 1: Relevant research studies on outbreak detection using machine learning approaches.

Research study	Year	Model	Disease type	Data sources	Objectives	Future work
Y. Wang et al. [18]	2020	PDE	Influenza flu	Twitter	To predict influenza or flu trend based on real-time data from OSM data.	In future work, this systematic approach will be extended to other types of an outbreak.
J. S. Coberly et al. [26]	2014	SVM	Dengue	Twitter	This study proposed a method that focused on geographic information about dengue fever. To associate the new case of dengue in a region with the reported cases by public health departments in the Philippines.	Overall tweet must be processed to achieve better insight into text data.
A. Alessa et al. [27]	2019	Linear regression	Flu, ILI	Twitter, CDC	This work aimed to develop an effective and accurate technique that efficiently utilizes OSM data to monitor the flu outbreak. To offer early detection, even for a novel epidemic.	There is a need for manual annotation to train the model for the entire OSM data.
K. Espina et al. [28]	2017	SVM, regression	Dengue, typhoid fever	Twitter	The purpose of the study is to strengthen the existing efforts to track disease outbreaks. This work has shown several dengue cases and typhoid fever in the Philippines to identify health-related tweets.	The work can be extended by exploring advanced machine learning techniques.
C. de a et al. [5]	2017	Linear regression	Dengue	Twitter	The key innovation of this work is to determine the importance of tweets regularly at the country and state level in Brazil for the fast detection and monitoring of a dengue outbreak.	Instead of traditional machine learning techniques, this work can be improved by utilizing deep learning techniques.
L. Sousa et al. [29]	2018	Naïve bayes, SVM	Dengue, chikungunya, zika	Twitter	To propose a VazaDengue system to detect mosquito-borne disease in tweets. To report and visualize new incidence of outbreaks.	In the future, there is a need to utilize Instagram content such as the classification of image data associated with the relevant post.
L. Chen et al. [19]	2016	Topic, hidden Markov model	Flu	Twitter	This work proposed syndromic surveillance of flu outbreak in tweets. To predict the flu outbreak, temporal topic models were deployed in this work.	In future work, the proposed state transition probabilities can be utilized for traditional epidemiological approaches.

deliver adequate statistics to speculate about its class when navigating the tree for the classification of a sample. Mathematically, it is formulated as follows:

$$(\mathbf{u},\mathbf{v}) = \left(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \ldots, \mathbf{x}_{\mathbf{n},\mathbf{v}}\right), \qquad (2)$$

where x shows the leaves in a tree, u represents the root node, and v is a subset in a tree.

6. K-Nearest Neighbor

One of the easiest and latest classification techniques is the K-Nearest Neighbor (KNN) method [29]. It can be considered as a simplified version of the Naïve Bayes classifier. The KNN approach does not need probability values to be considered, unlike the Naïve Bayes technique. "K" is the KNN algorithm that is known to take "poll" from the number of nearest neighbors. For the same sample item, the specification of distinct characteristics for "K" can produce separate classification accuracy. Mathematically, it is formulated as follows:

$$y = \frac{1}{K}\sum_{i=1}^{K} ui, \qquad (3)$$

where **ui** is the **ith** instance of the samples and **y** is the predicted result.

7. Support Vector Machine

Support Vector Machine is a classification technique used for supervised machine learning [33]. SVM operates by sorting the data into different classes by finding a line that is often referred to as a hyperplane that separates the set of data into categories. For text categorization, the fundamental concept behind linear SVM is to evaluate a hyperplane that separates the dataset or documents. The mathematical formulation of SVM is as follows:

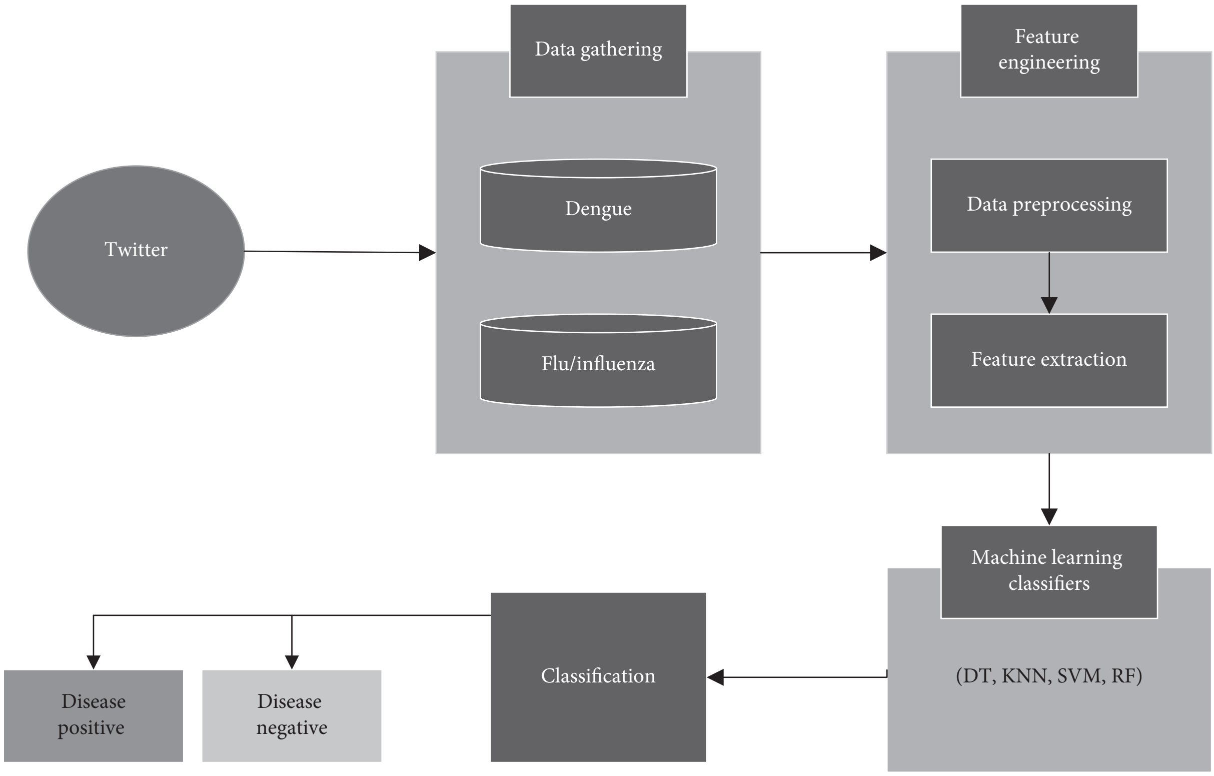

Figure 1: Proposed seasonal outbreaks classification methodology.

Table 2: Sample tweets related to the dataset for a training set.

Tweet#	Tweet	Class
1.	Day 1 after the Oxford-AstraZeneca vaccine and I have woken up with such a sore arm and feeling like I have got the flu.	Positive
2.	Flu season has been stopped in its tracks this winter.	Negative
3.	My daughter Francine has contracted dengue fever. She is in a critical point of the fever she is very weak and ill. Her chance of survival is reducing every day as her blood platelets have dropped dangerously.	Positive
4.	Dengue season is starting. Make sure you all are removing/changing water from coolers, indoor plants, and other small containers at least once a week.	Negative
5.	Flu season is in full swing, and we are currently in the usual peak months of December and February.	Negative
6.	May the healthcare system be able to handle the added burden of dengue season.	Negative
7.	Been fighting dengue for the last 5 days. Today I finally get to rest.	Positive
8.	My 20-year-old nephew just lost a close friend to dengue. Feeling a little numb and blank on hearing the news.	Positive

Table 3: The number of labelled tweets split over dengue and flu.

Disease type	Number of positive tweets	Number of negative tweets	Total
Dengue	1,290	2,210	3,500
Flu	900	1,600	2,500
Total	2,190	3,810	6,000

$$S = \{(x_1, y_1),\ (x_2, y_2),\ (x_3, y_3), \ldots\ldots\ldots\ldots\ldots\ldots,\ (x_n, y_n)\}. \tag{4}$$

The definition of the above equation is as follows: S represents sample of dataset and x will be associated with the y value showing if the item or features refer to the class.

8. Random Forest

Random Forest contains a huge amount of a DT that works as an ensemble [34, 35]. Each tree in an RF spits out a class prediction and the most voted class comes to be the prediction of the model. The basic idea about RF is that it is an ensemble

Table 4: The detailed results of the performance.

Algorithm	Precision (%)	Accuracy (%)	Recall (%)	F1 measure (%)
DT	86	84	87	86
KNN	87	85	85	86
SVM	87	85	85	84
RF	88	87	87	88

method and consists of several DTs that are close to the set of several trees in a forest [34]. The overfitting of the training data is also triggered by DTs, resulting in a high variance in the endpoint of classification for a minor alteration in the input features. They are very responsive to their training data, making them vulnerable to errors in the test dataset. Using the various sections of the training dataset, the various DTs of an RF are learned. To identify a new sample, on each DT of the forest, the input vector of that dataset must be passed down. A different part of the input vector is then considered by each DT and delivers a classification result. The forest then decides to classify the most "votes" for discrete classification results or numeric classification outcome for the aggregate of all forest trees. As the RF algorithm evaluates the effects of several different DTs, the variance resulting from considering a particular DT for the same dataset can be minimized. Mathematically, it is formulated as follows:

$$y = \frac{1}{Z}\sum_{z=1}^{z} yz\left(x'\right), \tag{5}$$

where z shows the training examples from x, y, and yz shows a regression or classification tree. After training, predictions for unseen samples x' can be made by averaging the predictions from all the individual regression trees on x'.

9. Performance Measures

Different measures are adopted for determining the performance of machine learning techniques [36]. The proposed approaches of the performance measures have multiple attributes and provide multiple results for determining the infected and the noninfected. For instance, some performance measures such as precision, accuracy, recall, F-score, and ROC curve are computed for determining the dengue and flu outbreak as infected or not infected (presented in Section 4). ROC curve is used for binary as well as multiclass classification tasks. When the ROC curve is better, the model would be to detect the target classes, where accuracy determines the ratio of correctly detected infected/noninfected cases in a class, and it is determined as

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{6}$$

where, TN: True Negative (correct refusal), that is, the dengue negative is categorized as not infected, TP: True Positive (correct detection), that is, the dengue positive is categorized as infected, FP: False Positive (type-I error), that is, the dengue negative is categorized as infected, FN: False Negative (type-II error), that is, the dengue positive is categorized as not infected.

Precision or positive predictive is the proportion between the expected appropriate disease positive cases and the total prediction of related and inappropriate disease positive is known as precision, and it can be computed as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{7}$$

A positive sensitivity or recall is the proportion of the actual infected obtained to the total number of actual infected, which can be measured as

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{8}$$

F-score or F1 measure can be determined with the help of harmonic mean. It organizes the data to better classify the test data and classify the findings obtained when a large number of incidents are not omitted.

$$\text{F} - \text{score} = 2\frac{1}{\text{recall}} + \frac{1}{\text{precision}}. \tag{9}$$

10. Results and Discussion

The selected machine learning algorithms (DT, KNN, SVM, and RF) were applied to the data and evaluated using precision, accuracy, recall, F1 measure, and ROC curve. To train the proposed model, we split the dataset into two parts using the test-train division method of scikit-learn: (i) Training data: we have utilized 80% of the data for training set through which the model learns and is used to fit the model. (ii) Test data: to evaluate the results of the model on the unseen data, we have applied 20% of the dataset as testing data.

Table 4 shows the detailed results. It is observed that using the RF model could lead to slightly improved results, while the DT and KNN also perform using confusion matrix for testing data with DT (Figure 2), KNN (Figure 3), SVM (Figure 4), and RF (Figure 5). The slight improvement in the results could be related to the length of the tweet summaries in the dataset. Similarly, Figures 6–9 demonstrate the ROC curve for testing dataset using DT (Figure 6), KNN (Figure 7), SVM (Figure 8), and RF (Figure 9). In the plots of confusion matrix and ROC curve, "1" shows positive class, while "0" shows negative class.

Figure 10 shows the training and test accuracy for each model (DT, KNN, and RF), while Figure 11 shows the performance accuracy achieved by adopting the selected machine

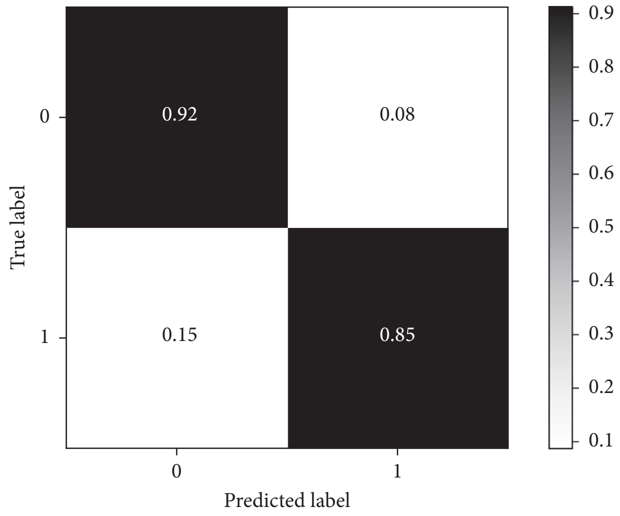

Figure 2: Confusion matrix for testing data using DT model.

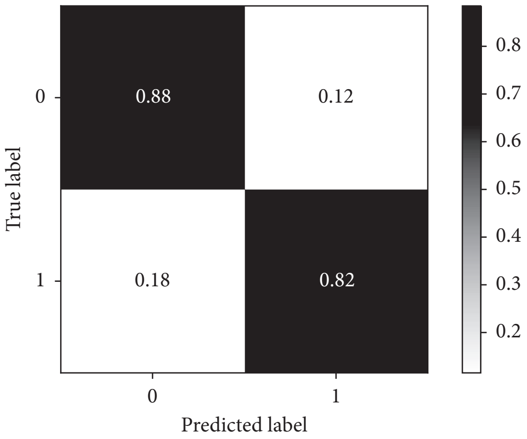

Figure 3: Confusion matrix for testing data using KNN model.

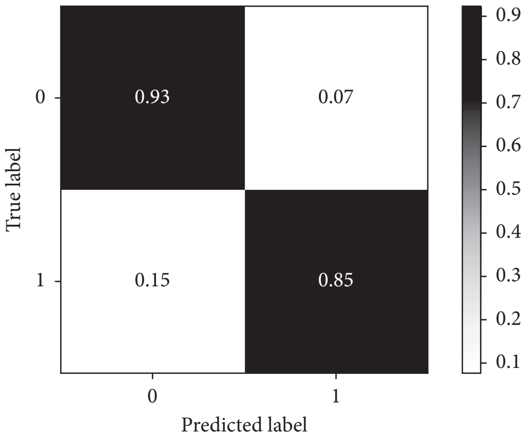

Figure 4: Confusion matrix for testing data using SVM model.

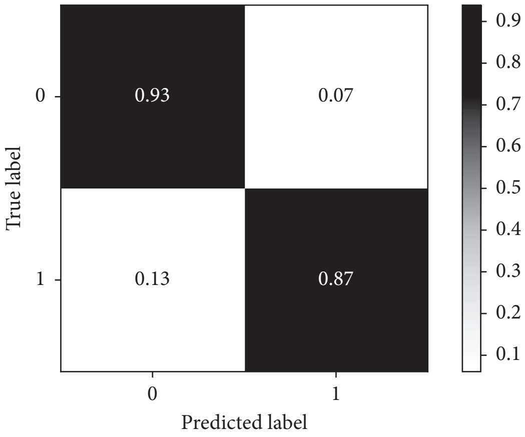

Figure 5: Confusion matrix for testing data using RF model.

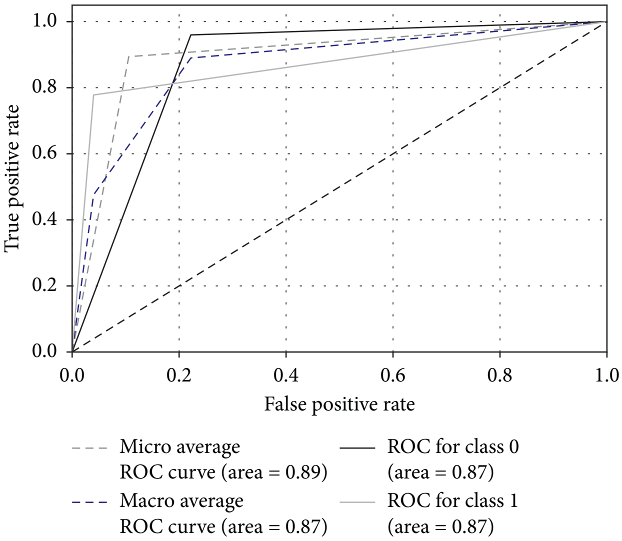

Figure 6: ROC curve for testing data using DT model.

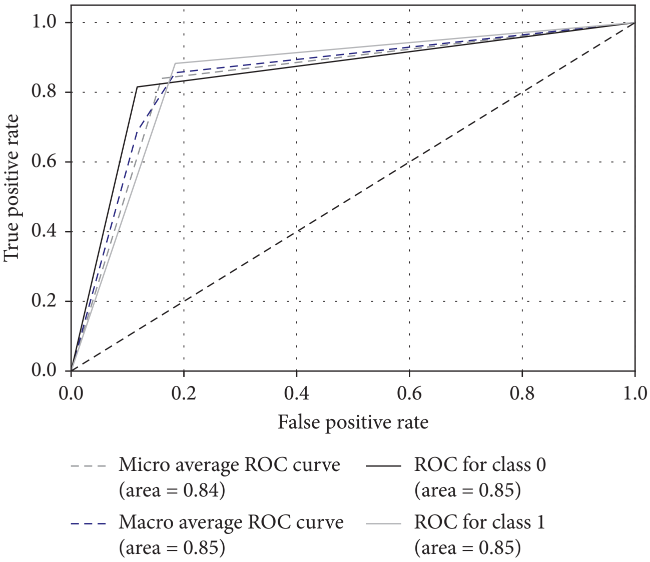

Figure 7: ROC curve for testing data using KNN model.

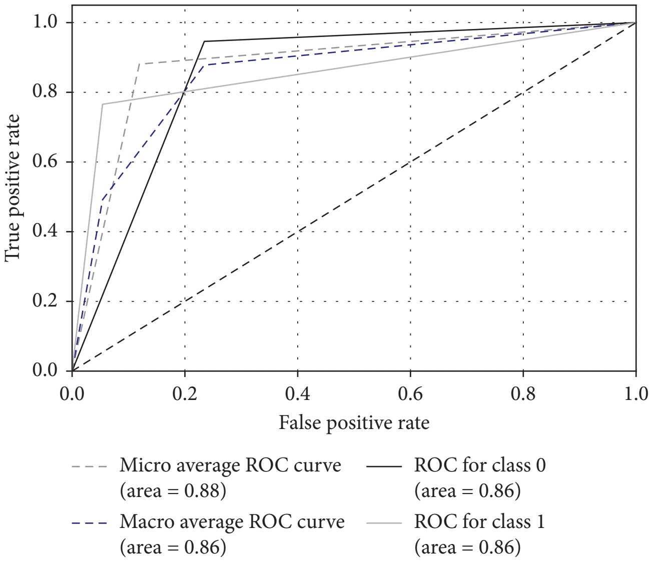

Figure 8: ROC curve for testing data using SVM model.

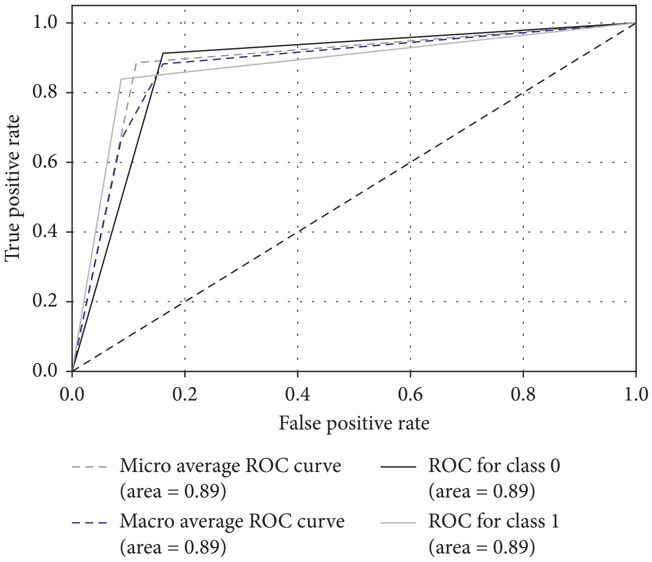

Figure 9: ROC curve for testing data using RF model.

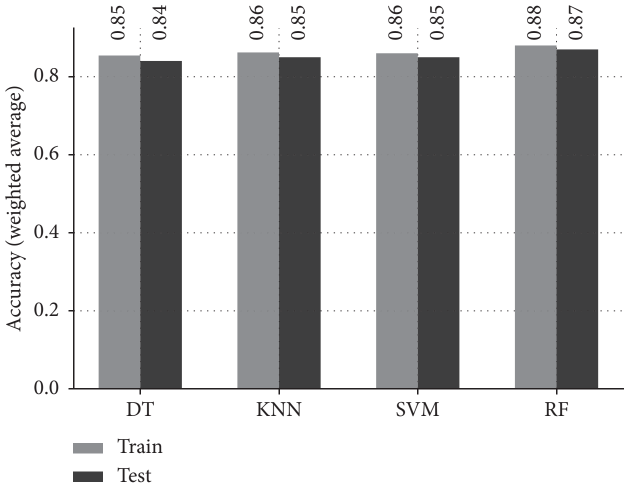

Figure 10: Results for training and test accuracy.

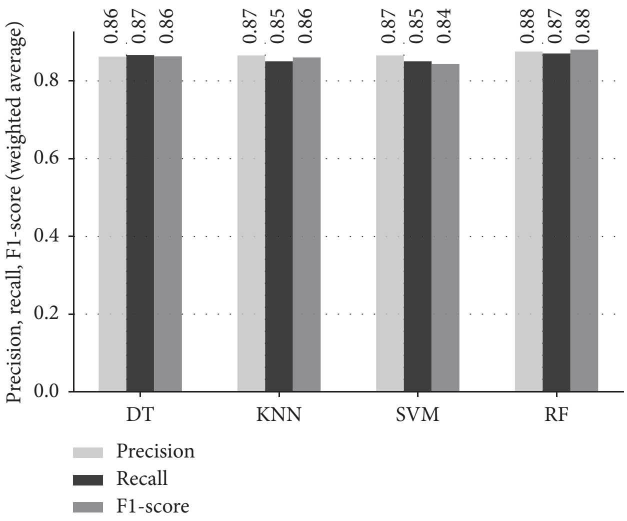

Figure 11: Results for precision, recall, and F1-score.

learning techniques DT, KNN, and RF in terms of different evaluation measures like precision, accuracy, and recall. Our evaluation proved that the performance of the proposed model is improved with a small amount of data, while it performed slightly better with RF in terms of all of the applied evaluation metrics, compared with DT, SVM, and KNN.

11. Conclusion

The important considerations of primary prevention and intervention provide identification and alert system of infectious disease outbreaks and modelling and evaluation of public health emergency. Recent years have seen a fast development of machine learning approaches in rapidly changing, dynamic, and data-rich environments to achieve these tasks. In this article, we summarized a set of current publications based on how to improve the use of the virtual world's sensor and OSM data to enhance seasonal outbreaks detection and early warning capacities. We also presented a collection of methods aimed at mapping the spread of seasonal outbreaks and estimating epidemiological data from Twitter messages.

This paper proposes a machine learning approach for the early detection of dengue and flu seasonal outbreaks. Four algorithms were applied (DT, KNN, SVM, and RF) and, for feature extraction, TF-IDF was used. The proposed methodology was evaluated on two datasets of 6000 labelled tweets on dengue and flu [10]. The results of the proposed method have been evaluated using confusion matrix performance evaluation techniques and ROC curve and their results are graphically visualized. Results showed that the RF classifier has outperformed SVM, DT, and KNN in terms of accuracy, precision, recall, and F1 measure. The proposed work offers favourable performance with total precision, accuracy, recall, and F1 measure ranging from 84% to 88% for conventional machine learning approaches.

Despite substantial findings revealed by the proposed model, it has certain drawbacks. Supervised learning was used in this research, because of which the data used for model training needed to be labelled. The model should be trained in an unsupervised way to avoid requiring labelled data. In the future, the proposed model may also be applicable as a surveillance system to quickly detect the transmission of coronavirus and COVID-19.

Data Availability

The data of this article are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research project was supported by a grant from the Research Center of the Female Scientific and Medical Colleges, Deanship of Scientific Research, King Saud University, Riyadh, Saudi Arabia.

References

[1] A. Alessa and M. Faezipour, "A review of influenza detection and prediction through social networking sites," *Theoretical Biology and Medical Modelling*, vol. 15, no. 1, pp. 1–27, 2018.

[2] E. Lau and P. Wark, "Twitter-Based influenza detection after flu peak via tweets with indirect Information : text mining study," *Journal of Medical Internet Research*, vol. 4, no. 3, pp. 1–27, 2019.

[3] H. Zaraket, N. Melhem, M. Malik, W. M. Khan, G. Dbaibo, and A. Abubakar, "Review of seasonal influenza vaccination in the Eastern Mediterranean Region: policies, use and barriers," *Journal of Infection and Public Health*, vol. 12, no. 4, pp. 472–478, 2019.

[4] S. Amin, M. I. Uddin, S. Hassan et al., "Recurrent neural networks with TF-IDF embedding technique for detection and classification in tweets of dengue disease," *Institute of Electrical and Electronics Engineers Access*, vol. 8, pp. 131522–131533, 2020.

[5] C. D. A. Marques-Toledo, C. M. Degener, L. Vinhal et al., "Dengue prediction by the web: tweets are a useful tool for estimating and forecasting Dengue at country and city level," *PLoS Neglected Tropical Diseases*, vol. 11, no. 7, pp. e0005729–20, 2017.

[6] I. D. Campbell, S. Lindskog, and A. I. White, "A study of the histidine residues of human carbonic anhydrase C using 270 MHz proton magnetic resonance," *Journal of Molecular Biology*, vol. 98, no. 3, pp. 597–614, 1975.

[7] H. Iso, S. Wakamiya, and E. Aramaki, "Forecasting word model: Twitter-based influenza surveillance and prediction," in *Proceedings of the COLING 2016*, pp. 76–86, Osaka, Japan, December 2016.

[8] S. Amin, M. Irfan Uddin, M. Ali Zeb, A. Abdulsalam Alarood, M. Mahmoud, and M. H. Alkinani, "Detecting information on the spread of dengue on twitter using artificial neural networks," *Computers, Materials & Continua*, vol. 67, no. 1, pp. 1317–1332, 2021.

[9] M. J. Paul and M. Dredze, "Social monitoring for public health," *Synthesis Lectures on Information Concepts, Retrieval, and Services*, vol. 9, no. 5, pp. 1–183, 2017.

[10] S. Amin, M. I. Uddin, M. A. Zeb, A. A. Alarood, M. Mahmoud, and M. H. Alkinani, "Detecting dengue/flu infections based on tweets using LSTM and word embedding," *Institute of Electrical and Electronics Engineers Access*, vol. 8, pp. 189054–189068, 2020.

[11] M. J. Paul, A. Sarker, J. S. Brownstein et al., "Social media mining for public health monitoring and surveillance," *Biocomputing*, vol. 2016, pp. 468–479, 2016.

[12] X. Zhang, L. Yang, Z. Ding, J. Song, Y. Zhai, and D. Zhang, "Sparse vector coding-based multi-carrier NOMA for inhome health networks," *Institute of Electrical and Electronics Engineers Journal on Selected Areas in Communications*, vol. 39, no. 2, p. 325, 2021.

[13] Z. Guo, Y. Shen, and A. K. Bashir, "Robust spammer detection using collaborative neural network in internet of thing applications," *Institute of Electrical and Electronics Engineers Internet of Things Journal*, vol. 2020, 1 page, 2020.

[14] K.-P. Yu, L. Tan, M. Aloqaily, H. Yang, and Y. Jararweh, "Blockchain-enhanced data sharing with traceable and direct revocation in IIoT," *Institute of Electrical and Electronics Engineers Transactions on Industrial Informatics*, vol. 2021, 1 page, 2021.

[15] K. Yu, L. Lin, M. Alazab, L. Tan, and B. Gu, "Deep learning-based traffic safety solution for a mixture of autonomous and manual vehicles in a 5G-enabled intelligent transportation system," *Institute of Electrical and Electronics Engineers Transactions on Intelligent Transportation Systems*, vol. 2020, 11 pages, 2020.

[16] K. Yu, L. Tan, X. Shang, J. Huang, G. Srivastava, and P. Chatterjee, "Efficient and privacy-preserving medical research Support platform against COVID-19: a blockchain-based approach," *Institute of Electrical and Electronics Engineers Consumer Electronics Magazine*, vol. 10, no. 2, pp. 111–120, 2021.

[17] C. Feng, K. Yu, M. Aloqaily, M. Alazab, Z. Lv, and S. Mumtaz, "Attribute-based encryption with parallel outsourced decryption for edge intelligent IoV," *Institute of Electrical and Electronics Engineers Transactions on Vehicular Technology*, vol. 69, no. 11, pp. 13784–13795, 2020.

[18] Y. Wang, K. Xu, Y. Kang, H. Wang, F. Wang, and A. Avram, "Regional influenza prediction with sampling twitter data and PDE model," *International Journal of Environmental Research and Public Health*, vol. 17, no. 3, p. 678, 2020.

[19] L. Chen, K. S. M. Tozammel Hossain, P. Butler, N. Ramakrishnan, and B. A. Prakash, "Syndromic surveillance of Flu on Twitter using weakly supervised temporal topic models," *Data Mining and Knowledge Discovery*, vol. 30, no. 3, pp. 681–710, 2016.

[20] M. J. Farrukh, L. C. Ming, S. T. R. Zaidi, and T. M. Khan, "Barriers and strategies to improve influenza vaccination in Pakistan," *Journal of Infection and Public Health*, vol. 10, no. 6, pp. 881–883, 2017.

[21] T. De Lima, R. Lana, T. De Senna Carneiro et al., "Dengueme: a tool for the modeling and simulation of dengue spatiotemporal dynamics," *International Journal of Environmental Research and Public Health*, vol. 13, no. 9, pp. 920-921, 2016.

[22] N. T. Toan, S. Rossi, G. Prisco, N. Nante, and S. Viviani, "Dengue epidemiology in selected endemic countries: factors influencing expansion factors as estimates of underreporting," *Tropical Medicine & International Health*, vol. 20, no. 7, pp. 840–863, 2015.

[23] P. Guo, T. Liu, Q. Zhang et al., "Developing a dengue forecast model using machine learning: a case study in China," *PLoS Neglected Tropical Diseases*, vol. 11, no. 10, pp. e0005973–22, 2017.

[24] Abdullah, S. Ali, M. Salman et al., "Dengue outbreaks in khyber pakhtunkhwa (KPK), Pakistan in 2017: an integrated disease surveillance and response system (IDSRS)-based report," *Polish Journal of Microbiology*, vol. 68, no. 1, pp. 115–119, 2019.

[25] F. A. Siregar, M. R. Abdu, J. Omar et al., "Social and environmental determinants of dengue infection risk in North Sumatera Province, Indonesia," *Asian Journal of Epidemiology*, vol. 8, no. 2, pp. 23–35, 2015.

[26] J. S. Coberly, C. R. Fink, E. Elbert et al., "Tweeting fever: can twitter be used to monitor the incidence of dengue-like illness in the Philippines?" *Johns Hopkins APL Technical Digest*, vol. 32, no. 4, pp. 714–725, 2014.

[27] A. Alessa and M. Faezipour, "Flu outbreak prediction using twitter posts classification and linear regression with historical centers for disease control and prevention reports: prediction framework study," *JMIR Public Health and Surveillance*, vol. 5, no. 2, pp. e12383–17, 2019.

[28] K. Espina, M. R. J. E. Estuar, and J. E. Estuar, "Infodemiology for syndromic surveillance of dengue and typhoid fever in the Philippines," *Procedia Computer Science*, vol. 121, no. 1, pp. 554–561, 2017.

[29] L. Sousa, R. De Mello, D. Cedrim et al., "VazaDengue: an information system for preventing and combating mosquito-borne diseases with social networks," *Information Systems*, vol. 75, pp. 26–42, 2018.

[30] Processing Raw Text, 2019, https://www.nltk.org/book/ch03.html.

[31] C. P. Medina and M. R. R. Ramon, "Using TF-IDF to determine word relevance in document queries," in *Proceedings of the first Instr. Conf. Mach. Lear.*, pp. 133–142, Piscataway, NJ USA, January 2003.

[32] V. F. Rodriguez-Galiano, B. Ghimire, J. Rogan, M. Chica-Olmo, and J. P. Rigol-Sanchez, "An assessment of the effectiveness of a random forest classifier for land-cover classification," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 67, no. 1, pp. 93–104, 2012.

[33] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *Journal of Machine Learning Research*, vol. 2, no. 11, pp. 45–66, 2001.

[34] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[35] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, “Comparing different supervised machine learning algorithms for disease prediction,” *BMC Medical Informatics and Decision Making*, vol. 19, no. 1, pp. 1–16, 2019.

[36] D. M. W. Powers, “Evaluation: From precision, recall and F-measure to roc, informedness, markedness & correlation,” *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.