IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 19, 2020, accepted July 7, 2020, date of publication July 13, 2020, date of current version July 28, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3009058

Recurrent Neural Networks With TF-IDF Embedding Technique for Detection and Classification in Tweets of Dengue Disease

SAMINA AMIN[1], M. IRFAN UDDIN[1], (Member, IEEE), SAIMA HASSAN[1], ATIF KHAN[2], NIDAL NASSER[3], (Senior Member, IEEE), ABDULLAH ALHARBI[4], AND HASHEM ALYAMI[5]

[1]Institute of Computing, Kohat University of Science and Technology, Kohat 26000, Pakistan
[2]Department of Computer Science, Islamia College Peshawar, Peshawar 25120, Pakistan
[3]College of Engineering, Alfaisal University, Riyadh 11533, Saudi Arabia
[4]Department of Information Technology, College of Computers and Information Technology, Taif University, Taif 21944, Saudi Arabia
[5]Department of Computer Science, College of Computers and Information Technology, Taif University, Taif 21944, Saudi Arabia

Corresponding author: M. Irfan Uddin (irfanuddin@kust.edu.pk)

ABSTRACT With the increased usage of Web 2.0 and data-affluent tools such as social media platforms and web blog services, the challenge of extracting public sentiment and disseminating personal health information has become more common than ever in the last decade. This paper proposes a novel model for Dengue disease detection based on social media posts alone. The model does not access any personal information of people or any medical record. The model extracts the presence of patients infected with Dengue disease based on tweets only and decides whether it is a general discussion about the disease, and no one is actually infected, or people are actually infected with that disease. The identification of people infected with Dengue is determined by clinical tests, but the propose technique is used for automatic surveillance and identification of regions where the spread is happening at an alarming rate and guide healthcare professional to take necessary actions to control the spread. This paper uses different machine/deep learning approaches to utilize tweets data for automatic and efficient disease detection. Experimental results demonstrate that the proposed model is able to achieve 92% accuracy compared to the current state-of-the-art techniques in this domain.

INDEX TERMS Deep learning, disease classification, machine learning, RNN, text processing.

I. INTRODUCTION

The increased usage of social media applications provides a good source of information to analyze users‘ feelings, opinions, and thoughts on multiple topics such as politics, sports, education, science, arts, etc. Social media users frequently post information or update status about their present circumstances. They may also share information about their daily life situations, or if there is an epidemic in a region that is rapidly growing or if they are infected by a disease. Analysis of real-time data from social media commonly termed as Social Media Analysis (SMA) has achieved considerable attention in recent years in the context of their analysis for detection of abnormal events/activities such as power outages, terrorism, assaults, disease detection [1], etc. For example, Doran *et al.* [2] developed a framework for detecting emergency events and public information about social, cultural, and political aspects using the potential of human sensors (i.e. tweets) to support smart city initiatives. Guntuku *et al.* [3] have presented a model that explored mental health using machine learning techniques to find out the pattern of depression among social media users in the US. Some other applications of SMA include sentiment analysis [4], [5] disaster prediction [6], earthquake [7], communication [8], [9], sports management [10], stock market fluctuations [11], political elections [12], [13] and healthcare [14], [15].

The rapid increase in Online Social Networks (OSNs) usage and level has led to a growing need for information extraction tools based on OSNs. A common application of SMA is promoting public health by identifying infectious diseases from social media posts to predict and monitor the epidemic outbreak [16]. Social media could effectively be used to identify disease infected people and impacts of

The associate editor coordinating the review of this manuscript and approving it for publication was Ting Li.

disease on health promotion (e.g., cancer, swine flu, depression, and dengue, etc.) with an intervention to promote public health [17], [18]. With the usage of SMA, we can detect patterns of early warnings about the disease and can decrease the time that passes between onset and detection. This was previously dependent on reporting cases of the disease by physicians and healthcare professionals [19]. The traditional methods of detecting epidemic are when people are infected with disease and report to the local health center, who can then inform related health care professionals/organizations to react and provide resources to control that epidemic. This process sometimes takes days or weeks when the data become available, and in some cases, precious lives are lost before a necessary action is taken.

Social media has been considered as a data source for tracking diseases. The content of social media provides useful information regarding the surveillance of early disease outbreaks. Among OSNs/SMA, Twitter captures awareness for public health monitoring and surveillance purpose since Twitter messages can be effectively retrieved [20]. Automatic surveillance based on social media [1] can track trends in the epidemic outbreak in real-time to predict, monitor, and minimize the risk caused by outbreak events. Early detection of an epidemic outbreak is essential for healthcare professionals to generate a reaction more quickly and efficiently. The epidemics outbreaks of infectious diseases such as Dengue [16], [21], [22] can cause death when those diseases epidemiologically break in a region. Dengue fever can resemble the flu, but it can become life threatening. Dengue infection is a mosquito-borne virus causing serious influenza-like-illness (ILI) and often causing a possibly fatal risk factor called severe Dengue fever infection. Dengue is one of the world's fastest proliferating infectious diseases. The provision of real-time surveillance, early warnings, and detection of infectious diseases regarding influenza or Dengue outbreak is therefore essentials for public health [23]–[25]. A large number of the health care professional and health organizations have now transformed their concentration on mining useful information from OSNs to better and quickly understand about the infectious disease outbreak and infected people in a region, to fill the gap between interaction and disease-infected people and discuss new research areas to detect early warnings and provide necessary resources.

Moreover, early warnings of disease detection can decrease the influence of seasonal epidemic outbreaks (i.e., Dengue or flu) in public health. SMA/OSNs can be used for disease surveillance to monitor the rate of epidemic outbreaks quicker than health care professionals and health organizations like the American Center of Disease Control and Prevention (CDC) [23]. CDC uses Influenza-like-Illness Surveillance Network (ILINet), a program used by health care professionals, to monitor early warnings of influenza outbreaks. Although it is a reliable method but costly and slow as it takes days or weeks when data becomes available. Numerous studies, therefore, focus on proposing solutions using SMA to monitor ILI and detect early warnings about epidemic outbreaks to perform real-time analysis. Social media platforms such as Twitter can be used to detect epidemic outbreaks among the public and can assist early warnings [26]. Through SMA, health care professionals/organizations can be alerted to provide necessary resources to control an epidemic.

Different machine learning approaches are introduced to detect epidemics. For instance, SVM, KNN, Logistic Regression (LR), etc. are used to detect tweets about diseases in literature. However, existing studies on disease detection are limited to the frequency of tweets about the disease. The more tweets from a region are about a disease, the more chances are that the region is infected with that disease. However, the frequency of tweets is not a reliable source for detecting epidemics. There are limited studies in the literature about the sentiment of tweets to identify infected people from diseases and the impacts of epidemics on a region at an alarming rate. In order to overcome limitations in the previous studies the main contribution of this work is to utilize Recurrent Neural Network (RNN) [27] with Term Frequency — Inverse Document Frequency (TF-IDF) [28] and n-gram [29]. However, there are many other embedding techniques such as Word2Vec [30] and Glove [31], etc. In this paper, different machine learning techniques are explored to utilize tweets data for automatic and efficient disease detection. The paper explores that RNN [27] is the most efficient machine learning technique to process the flow of sequence data and efficiently detect the number of infected people in tweets.

The main objectives of this paper are (1) Mining the microblogging data to understand the sentiment of tweets regarding the Dengue disease. A collection of tweet dataset is collected from September 2017 to November 2019. A sample/subset of these tweets is manually annotated from April 2018 to December 2019, with the assistance of three human annotators and the recommended annotations are eventually acknowledged with the inter-annotator agreement (described later). We train machine learning models such as Artificial Neural Networks (ANNs) and Long-Short Term Memory (LSTM) on the training dataset. The trained model is used to detect and classify the infected people; (2) Proposing RNN method for Dengue disease detection with word embedding techniques: TF-IDF [28] and n-gram [29] to capture the broader context of the words in social media text for better classification. The results demonstrate that the performance of the proposed solution is improved compared to other state-of-the-art algorithms, such as LR, SVM, and Naïve Bayes (NB).

The rest of the paper is organized as follows. Section 2 gives a brief overview of the background studies on the existing solutions that detect epidemic outbreaks using machine learning techniques. Section 3 presents a detailed methodology of the proposed work. However, the algorithms that are utilized for the analysis of tweets are also demonstrated in section 3. Section 4 presents the results and discussions. Section 5 concludes the paper and provides directions to the future research trend.

TABLE 1. The Dengue incidence reported at country level.

Citation	Year	Country	# +ve Incidence	Founded
[42]	2018	Pakistan	74,595	1982-2014
[43]	2018	Pakistan	47,836	1994-2016
[44]	2017	China	52,749	2009-2014
[45]	2018	India	683,545	2009-2017
[24]	2017	Brazil	106,558	2012-2016
[46]	2016	Sri Lanka	125,000	2012-2014
[47]	2019	Bangladesh	40,476	2000-2017
[48]	2019	Philippine	842,867	Annual average
[49]	2017	Global Estimation	50-200M	Annual average

II. RELATED WORK

Multiple studies have been conducted in the field of SMA for detecting the flu outbreaks using machine learning methods to filter irrelevant tweets. Xue *et al.* [32] developed a Support Vector Regression (SVR) based approach which focused to predict influenza regional based rates in the US. In this work, two datasets were used: one that supports US-based twitter data and second CDC data to optimize the parameters of SVR. In order to slow the spread of H1N1 influenza, the CDC has used twitter platform for posting instructions to prevent flu. The characteristics of epidemic outbreaks are highly dynamic with temporal and spatial aspects in social media. Many researchers have applied content analysis and machine learning methods such as SVM, Linear or LR, KNN, etc. and have achieved an accuracy of approximately around 88% [33]. The goal of their research was to monitor public concerns about diseases by classifying tweets into disease symptoms and non-disease symptoms. Cambria *et al.* [14] analyzed the public sentiment in the US with respect to swine flu and H1N1 by utilizing regression models and statistical methods. Similarly, Alessa and Faezipour [34] have described the potential of social media posts to detect disease outbreaks and provide early warnings to monitor the epidemic. The target of their work was to detect flu using classification, and prediction of the flu outbreak to evaluate linear regression. Cambria *et al.* [14], proposed Artificial Intelligence (AI) and semantic based approach that was deployed for identifying and analyzing the sentiment of patients in Natural Language Processing (NLP) text on a Web ontology. Lee *et al.* [35] focused on the frequency of disease in tweets, where machine learning approaches were utilized for detecting disease tasks in which two diseases i.e. flu and cancer were identified based on the frequency. However, the frequency of tweets about a disease does not give any information about the possibility of disease-infected people.

Dengue outbreak is a viral infection that is spreading worldwide [21]. The recorded Dengue cases across the most Dengue-affected countries are presented in Table 1. This table demonstrates that there is a need to find AI-based solutions to look for the prevention of Dengue. Data on Dengue monitoring are highly required to better detect the Dengue fever outbreak and to assess the effect on preventive intervention [36]. Another interesting work carried out by Missier *et al.* [22], where a model was presented to track Dengue epidemics in tweets. Their model has been evaluated on 1,000 tweets; machine learning models such as NB and LDA-based topic modeling were used for analysis. The 2017 Dengue outbreak has been analyzed in the Philippines [37]. In order to classify health related tweets, this work has shown a range of Dengue cases and typhoid fever in the Philippine using SVM for classification and regression model for possible disease incidence.

It has been observed in the literature that previous studies on disease detection from SMA are based on traditional machine learning methods. In addition, the important features like disease-infected people, sentiment about the disease, and impacts of epidemic on a region at an alarming situation are not considered. However, Wang *et al.* [38] used RNN for the prediction of patients from eight different diseases using NTCIR13-MedWeb dataset and 1,920 tweets for model evaluation. Some other relevant studies can be found in [39]–[41].

To the best of our knowledge, there is no benchmark dataset available on the disease of Dengue that provides observations of public sentiment. We have extracted the sentiment of a tweet about people infected with Dengue or flu. In order to track the epidemic, this approach will help health analyst to see how a disease spreads as they have real-time information from social sensors (referred to as tweets) where a disease infected person is increasing at an alarming rate in social media to alert health care workers. This work is a collaboration among AI, health care analysts, and social media text analysis to know more about social media information to detect epidemic outbreaks in social media text. This work overcomes limitations in the previous studies, by utilizing RNN with TF-IDF embedding and a large number of tweets.

III. METHOD AND SYSTEM DESIGN

In this section, we explain the architecture of the proposed model that addresses the issue of detecting the number of people infected with the Dengue disease by utilizing machine

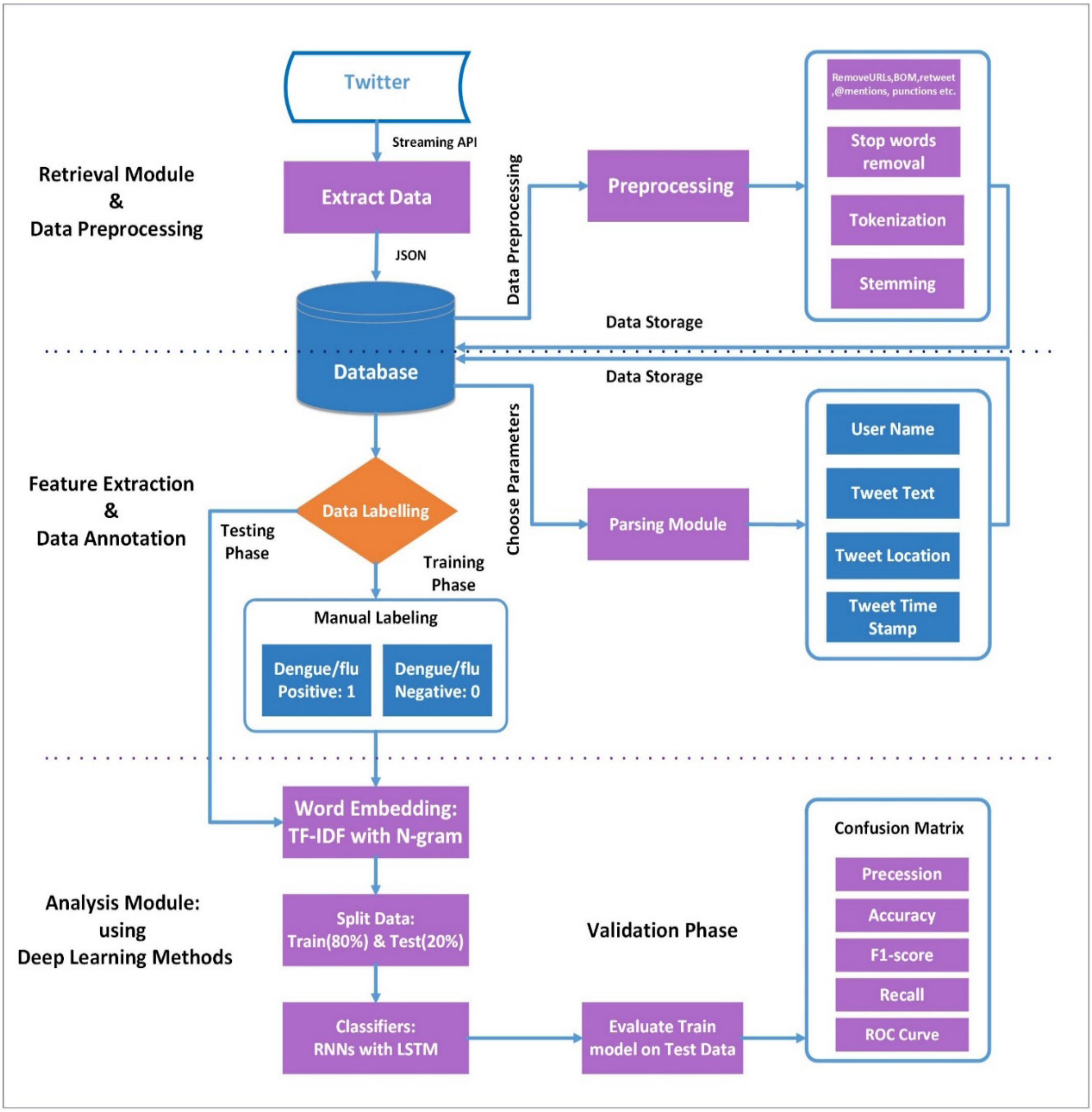


FIGURE 1. Proposed methodology adopted for disease classification.

and deep learning approaches such as Deep Neural Network (DNN), RNN with LSTM and word embedding techniques. The proposed framework consists of four modules to analyze given tweets data: 1) Data collection and preprocessing module; 2) Data modeling (Feature selection) module; 3) Classification module; and 4) Evaluation module. The different modules of the proposed method are presented in the framework shown in Figure. 1. The details of each module are explained in the following subsections.

A. DATA COLLECTION

The focus of the proposed model is to detect infected people with Dengue disease by analyzing tweets obtained from social media. In this work, Twitter Streaming API (Application Programming Interface)[1] is utilized to collect tweets on Dengue and flu. Twitter Streaming API is an open source interface that allows researchers to access the tweets in real-time in JavaScript Object Notation (JSON) format. Using this scraper, 359,410 tweets are collected from September 2017 to November 2019, having keywords #Dengue, #Denguefever or #Denguevirus. The other most commonly encountered hashtags founded in the corpus of this paper are displayed through the pie chart in Figure 2.

[1] https://developer.twitter.com/

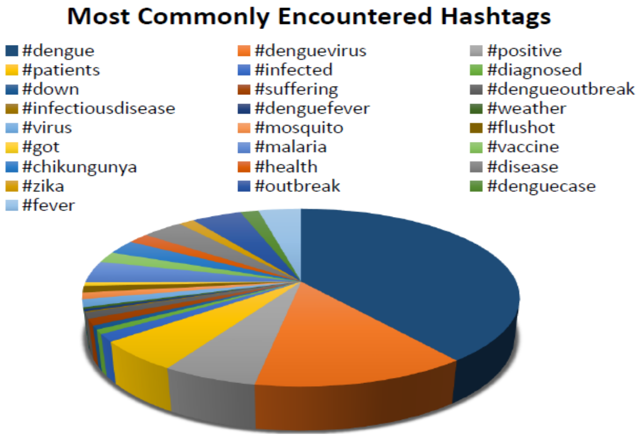


FIGURE 2. Most commonly encountered hashtags in dengue corpus.

B. DATA ANNOTATION AND ANNOTATORS AGREEMENT LEVEL

Tweets are labeled manually. They are labelled based on the information whether a tweet contains someone infected with Dengue or not. This labelled data is used to train the model. In total there are 359,410 tweets; however, a subset of 6001 tweets of the corpus from April 2018 to October 2019 is used to train the model. They are annotated with the assistance of three annotators. Annotators are employed from different fields of life to eliminate biases in labelling.

The manual labelling of social media dataset has been in action with the assistance of human specialists from multiple domains to build numerous baseline dataset [50]–[52], but it is costly and time-consuming. The dataset was annotated manually through human annotators. At the tweet level (sentence level), the labelling was conducted, and the specified annotations have been acknowledged by utilizing inter-annotator agreement. Annotation is the procedure of labelling data to train the model to make it easier to learn and perform efficiently. In the inter-annotator agreement, the measurement of more than two annotators indicates how efficiently the labels are calculated and how efficiently they take the same judgment to annotate a certain label throughout the corpus [53]. The usual way to solve the problem of agreement among annotators, Kappa test proposed by McHugh [54] for agreement with the help of Raters/Annotators being utilized. It can be observed that the stronger the agreement, the more probable it is that a reliable annotation method can be built. The calculation of Cohen's Kappa [54], was applied in this paper and it was found that the measurement of the inter-annotation agreement of three annotators is (Kappa = 0.853). Following the agreement level presented by Fleiss *et al.* [55], there is a strong agreement (Kappa = 0.853) between the user-based annotations.

With the help of multiple annotators, the process of annotating the data is as follows. Each tweet is annotated as either '1' or '0'. That is to say, if the tweet is related to Dengue positive instance, a label '1' is assigned. If the tweet is Dengue negative that is if it only contains the word Dengue, a label '0' is assigned. For simplicity, some example tweets are as follows: a) Dengue Positive Tweet "*My friend and his sister are both diagnosed with Dengue and are in urgent need of O+ blood tomorrow by* 8 *am. I request you all to pray for their good health*" (label =1), and b) Dengue Negative Tweet "*Today was a world day of Dengue. I also gave a lecture on Dengue*" (label = 0). The preprocessed and annotated dataset (6001 tweets), distributed between two classes of Dengue positive and Dengue negative and acknowledged via inter-annotator agreement is shown in Table 2.

TABLE 2. Manually annotated data in the corpus distributed over Dengue positive and negative tweets by the annotators from multiple domains and the annotations are acknowledged trough inter-annotator agreement.

Annotation Order	# Dengue +ve	#Dengue -ve	Total
Manual Annotation	2190	3811	6001

C. DATA PREPROCESSING

The proposed framework retrieves all tweets that contain the word "Dengue", #Denguefever or #Denguevirus, and related information from Twitter using Twitter Streaming API and stores it in a database. Related information is; tweet text, tweet location, tweet time stamp, and username.

Extracting useful structured representations of text from a disorganized corpus of noisy text is a challenging problem. Tweets are short and self-contained and are, therefore, not composed of complex discourse structure as is the case for texts containing narratives. Due to the 140-character limit of tweets text and its informal nature, the posted tweets are very noisy in nature. People tend to use abbreviations and emoticons, especially hashtags, as well as less distinctive means as sarcasm and humor. Humans can easily capture the meaning of a tweet, but the same task has not always been the same for the computer. This step aims to present data in a way, which can be effectively analyzed and to improve their performance by removing the posts that are irrelevant to the corpus.

Once the raw tweets are retrieved, only the tweets written in English are kept, the remaining tweets (belonging to any other language) are removed. After that, data is processed using preprocessing methods NLP to optimize the text to be used as parameters/features. This includes stop words removal, stemming, and tokenization. Short tweets containing less than four words are eliminated, as it is not possible to retrieve sentiment information from short tweets [56]. Regular expressions are utilized to eliminate punctuations marks, hyperlinks, emoticons, special characters, URLs, retweets and @mentions, etc.

In NLP, stop words are the words that are too common to operate as a helpful function/feature. These words do not transmit any semantic to the text or sentences in which they occur. For instance, in almost all documents, the article 'the' and the verbs 'is', 'be', 'are' and prepositions 'of', 'to', 'at' etc. exist. To eliminate such stop words from the documents, NLTK (Natural Language Tool Kit) packages in Python [57] are used.

Once the irrelevant elements are filtered out from the tweets, it is then tokenized into word using the NLP tokenization technique in Python. Tokenization is the process of splitting the sequence of text into word/components. Examples of such tokens or components are; words, keywords, statements, or even whole sentences, etc. In text mining techniques such as NLP and Information Retrieval (IR), the very common step of preprocessing is stemming known as word normalization approaches. Stemming is the process to reduce the inflected word to its root word stem such as liking, likely, liked, likes, etc. can be combined in the root word "like". After tokenization, each word of the corpus is then transformed to stemming (stem or root word) using the NLTK porter stammer technique in Python.[2]

D. FEATURE SELECTION

In NLP, one of the most important processes is feature selection. In the feature selection process, the relevant and most useful features are retrieved from the corpus to improve the efficiency and performance of the model. The feature converts the text into a vector space that has a number of features. However, in this paper, we have used TF-IDF and n-gram models to organize data for evaluation. Textual data needs to be converted to numbers and the most widely used method to process textual data into numbers is TF-IDF [28]. In TF-IDF, text data is transformed into vectors without compelling the exact sequence of word order into consideration. Each word in the corpus is correlated with a number by TF-IDF that shows how significant each word to the corpus. Once the words are converted into numbers, the numerical values of TF-IDF are fed to supervised learning classifiers [58] in a context that machine learning methods can interpret.

Another embedding technique utilized in this paper is the n-gram embedding technique. n-gram is an efficient technique as it offers the sequence of words in a corpus [29]. In order to convert text data into a weighted vector and to allocate probabilities efficiently to a sequence of a word in a tweet, the n-gram model is utilized. For the purpose to understand the n-gram model, consider an example tweet such as, "*My sister got infected with Dengue hope she will recover soon.*" The n-gram interpretation for 3-gram (N – 1 = 2, in this case, it predicts the occurrence of a word based on its previous two words) representation will convert the example tweet given above as, "*my sister got*", "*sister got infected*", "*got infected with*", " *infected with Dengue*", "*with Dengue hope*", "*Dengue hope she*", "*hope she will*", "*she will recover*", "*will recover soon*". There are other n-gram models such as: unigram (1-gram) model consists only one word and the bigram (2-gram) model consists of two words as it determines the appearance of a word given in a sentence only its subsequent word (N – 1 = 1).

E. CLASSIFICATION IN DENGUE

With the growing availability of social media, the proposed work develops a novel approach for disease classification based on social media posts to detect the Dengue outbreak by utilizing machine learning classifier methods. Once the preprocessed features are retrieved, the classifier techniques can be utilized to build the model for disease classification. After the detailed study of related work, in our research, we planned to use machine learning approaches (LR, SVM, and NB) as a baseline model to deep learning approaches (NN, DNN, and LSTM) to classify disease infected tweets. The details of each classifier are demonstrated in the following subsections.

1) MACHINE LEARNING CLASSIFIERS

a: LOGISTIC REGRESSION (LR)

In machine learning, the most popular classification method is LR [59]. It determines the categorical variable based on a set of independent variables. Examples of such variables include 'spam/not spam', 'yes/no', 'pass/fail', and 'positive/negative' (i.e., in our case 'disease infected/not-infected').

Like linear regression, in LR the straight line is not explicitly fitted to the observations (i.e., data). Alternatively, the observations fitted with the S shaped curve called sigmoid. In addition, to utilize the logistic sigmoid function the probability, P, is estimated by calculating the correlation between dependent categorical variable (prediction/output) and one or even more independent variable (features matrix) that is depicted in Eq.1, where x represents input features, w shows weight and b is a bias value. In this work, the LR takes continuous input features (i.e., the amount of words in a tweet) and generates output to a discrete number such as 0 or 1, where 0 indicates disease negative class (not infected) and 1 indicates disease positive (infected people with Dengue).

$$\begin{aligned} \mathrm{P} &= \frac{\mathrm{e}^{\mathrm{x}}}{\mathrm{e}^{\mathrm{x}}+\mathrm{e}^{-\mathrm{x}}}, \\ \text{with } \mathrm{z} &= \mathrm{b}+\mathrm{w}_1\mathrm{x}_1+\mathrm{w}_2\mathrm{x}_2+\ldots+\mathrm{w}_\mathrm{n}x_\mathrm{n} \end{aligned} \tag{1}$$

b: SUPPORT VECTOR MACHINE (SVM)

SVM is a non-probabilistic supervised learning model [60]. It provides a straight-line boundary between vectors relating (fitting) to a specific class (group/category) and vectors that do not fit that class and can be seen mathematically in Eq.2. However, a hyperplane is a path where the input data is classified into two classes (Dengue positive and dengue negative in our case), where w is a weight and X is an input matrix of all features. Per instance features are represented by $\mathrm{x}_1 \ldots \mathrm{x}_\mathrm{n}$. Y is the output result. SVM can be extended towards any type of vectors that computes or encodes any type of data. In order to classify text data to utilize the efficiency of SVM, the texts have to be transformed into vectors. In order to trace the appropriate hyperplane that distinguishes the gap between two classes such as one for the vectors fitting to the related class and one for the vectors not related to that class

[2] https://www.nltk.org/_modules/nltk/stem/porter.html

(i.e., in our case disease positive and negative class).

$$\text{Hyperplane}: \mathrm{w}^{\mathrm{T}}\mathrm{x} = 0, \text{ and}$$
$$\text{Line}: \mathrm{Y} = \mathrm{w}_1\mathrm{x}_1 + \mathrm{w}_2\mathrm{x}_2 + \ldots + \mathrm{w}_\mathrm{n}\mathrm{x}_\mathrm{n} + \mathrm{b} \quad (2)$$

c: NAÏVE BAYES (NB)

Unlike SVM, NB is a probabilistic supervised learning model [61], based on Bayes theorem. It is commonly used in binary classification problems (email spam/not, yes/no, fail/pass, etc.), sentiment analysis, and recommender system, etc. In order to classify text data, the fundamental concept of the NB method is to calculate the probabilities of categories allocated to the corpus, tweet text known as features, by estimating the joint probability of classes and text (i.e., words in a tweet). It is depicted mathematically in Eq.3, where P(c|x) is a conditional probability, while P(c) is the prior probability and P(x) is the probability of observing x.

$$\mathrm{P(c|x)} = \text{with}$$
$$\mathrm{P(c|X)} = \mathrm{P(x_1|c)xP(x_2|c)x\ldots xP(x_n|c)xP(c)} \quad (3)$$

This method calculates the probability of a tweet being infected with Dengue or not. For this purpose, the parameters (or features) of a tweet, particularly the number of words in a tweet are considered. This means that NB methods operate by associating the use of features (i.e., words), with infected and non-infected tweets, and then computing a probability that a tweet demonstrates that a person is or is not infected with Dengue using Bayes' theorem. First, the text data is transformed into vectors by utilizing the embedding techniques, and then the values of the vectors are fed into the NB classifier. The efficiency of the model is evaluated on the test phase, which is not considered during the training phase of the method. Finally, in order to get a better insight of the model, the 10-Fold cross validation, performance metric such as precession, recall, and F1-Score, etc., is also measured.

2) DEEP LEARNING CLASSIFIERS

a: ANN/DNN

ANN is a mathematical based method typically used for classification and prediction purposes for both numerical and categorical data [62], [63]. The general structure of ANN consists of one input layer, one or more hidden layers, and one output layer. An ANN that has more than one hidden layer is known as DNN. Each layer consists of multiple nodes and the number of nodes in each layer be influenced by the number of parameters (features) in data and data type. Each layer has assigned separate weights. then these weights are combined with the input features and then moved to the next layer. ANN learns optimized weight values on each layer by back-propagating the estimated or predicted error (true output – estimated output) to the preceding layers. The back-propagation approach is aimed to increase the model accuracy and decreases the estimation error so that the optimal output can be obtained for specific input features on completion of the training. The output at each layer is generated by activation function that can be rectified linear unit (ReLU) or sigmoid. Where ReLU is assigned at each layer and provides the output ranges between 0 to max [0, ∞), while sigmoid is allotted at the output layer that generates the probability as output between 0 and 1. The pseudocode for ANN and DNN is depicted in algorithms 1 and 2, respectively.

Algorithm 1 Pseudocode for ANN

```
Begin
Input: Tweets corpus
Output:  tweet infected with Dengue or not
$W^1$, $W^2$, $b^1$, $b^2$ $\leftarrow$ random initializer
while i $\leq$ 500 do
    $Z^1 \leftarrow$ np.dot ($W^1$, X) $+b^1$
    $A^1 \leftarrow$ ReLU($Z^1$),where ReLU (Z) $=$ max (0, z)
    $Z^2 \leftarrow$ np.dot ($W^2$, $A^1$)$+ b^2$
    $A^2 \leftarrow$ sigmoid ($Z^2$),// sigmoid
    $Z^2 = \frac{1}{1+e^{-z}}$
    $L(A^{2,}, Y) \leftarrow -Y.\log(p(Y)) + (1 - Y).\log(1 - p(Y))$
    $W^1 \leftarrow W^1 - \propto \frac{\partial L(A2,Y)}{\partial W^1}$
    $W^2 \leftarrow W^2 - \propto \frac{\partial L(A2,Y)}{\partial W^2}$
    $b^1 \leftarrow b^1 - \propto \frac{\partial L(A2,Y)}{\partial b^1}$
    $b^2 \leftarrow b^2 - \propto \frac{\partial L(A2,Y)}{\partial b^2}$
end while
End
```

In the first step, data is converted to TF-IDF as discussed above. There are around 10,000 TF-IDF values (referred to as features) for the whole train dataset. in the training phase, the data (i.e., TF-IDF values of text) is processed by ANN and can be seen in algorithm 1. After the training, the values of W (weight) and b (bias) are optimized using the gradient descent algorithm to find the minimum loss value. We can then use these values of W and b to make predictions.

b: RNN AND LSTM

ANN is typically used for classification purposes [62]; however, there are some limitations when the sequence data such as words or sounds are processed. For instance, in ANN input and output features are fixed in size, which may not be the case when sentences are processed. In this paper, we are processing text in tweets and, therefore, the size of all input is not the same (i.e., one sentence may have 5 words, another may have 8 words and so on). Therefore, ANN may not be able to produce efficient results. Moreover, ANN does not share features learned across different positions in texts. In order to make a good prediction of a sequence of words, learned features must be shared across different positions. RNN is designed to overcome these limitations in ANN. A simplified notation of RNN is shown in Eq. (4), where $a^{<t>}$ is an activation function, which is typically tanh or ReLU.

Algorithm 2 Pseudocode for DNN

```
Begin
Input: Tweets corpus
Output: tweet infected with Dengue or not
$W^1$, $W^2$, $W^3$, $W^4$, $b^1$, $b^2$, $b^3$,$b^4$ ← random initializer
while i ≤ 500 do
    $Z^1$ ← np.dot ($W^1$, X) +$b^1$
    $A^1$ ← ReLU($Z^1$),where ReLU (Z) = max (0, z)
    $Z^2$ ← np.dot ($W^2$, X) +$b^2$
    $A^2$ ← ReLU($Z^2$)
    $Z^3$ ← np.dot ($W^3$, X) +$b^3$
    $A^3$ ← ReLU($Z^3$)
    $Z^4$ ← np.dot ($W^4$, $A^3$)+ $b^2$
    $A^5$ ← sigmoid ($Z^5$), //sigmoid ($Z^2$) = $\frac{1}{1+e^{-z}}$
  $L(A^3, Y) - Y.\log(p(Y)) + (1 - Y).\log(1 - p(Y))$
    $W^1 \leftarrow W^1 - \propto \frac{\partial L(A2,Y)}{\partial W^1}$
    $W^2 \leftarrow W^2 - \propto \frac{\partial L(A2,Y)}{\partial W^2}$
    $b^1 \leftarrow b^1 - \propto \frac{\partial L(A2,Y)}{\partial b^1}$
    $b^2 \leftarrow b^2 - \propto \frac{\partial L(A2,Y)}{\partial b^2}$
    end while
End
```

The input to this activation function is the output from the previous unit ($a^{<t-1>}$) multiplied by its weight (w_a), added with the input for the current unit ($x^{<t>}$) multiplied by its weight (w_x). A bias term (b_a) is also added. y ^^(<t>) is the probability of output, computed using the output activation function that can be sigmoid in case of binary classification and softmax in case of multiclass classification. More details of RNN are given in [27].

$$\begin{aligned} a^{<t>} &= g\left(W_a a^{<t-1>} + W_x x^{<t>} + b_a\right) \\ \hat{y}^{<t>} &= g\left(W_y a^{<t>} + b_y\right) \end{aligned} \tag{4}$$

A typical problem in RNN is long term dependencies [64] when processing a long sentence. Solutions in the form of LSTM [65] and GRU [66] are proposed to avoid long term dependencies. In this paper, we have utilized the *LSTM* technique as it has the ability to preserve a memory cell ($\tilde{c}^{<t>}$) at timestamp (t). The output ($a^{<t>}$) is calculated using the following Eq. (5) [65].

$$\begin{aligned} \tilde{c}^{<t>} &= \tanh\left(W_c\left[a^{<t-1>}, x^{<t>}\right] + b_c\right) \\ \Gamma_u &= \sigma\left(W_u\left[a^{<t-1>}, x^{<t>}\right] + b_u\right) \\ \Gamma_f &= \sigma\left(W_f\left[a^{<t-1>}, x^{<t>}\right] + b_f\right) \\ c^{<t>} &= \Gamma_f * c^{<t>-1} + \tilde{c}^{<t>} * \Gamma_u \\ \Gamma_o &= \sigma\left(W_o\left[a^{<t-1>}, x^{<t>}\right] + b_o\right) \\ a^{<t>} &= \Gamma_o * c^{<t>} \ldots\ldots\ldots\ldots\ldots\ldots\ldots \end{aligned} \tag{5}$$

Consider the above Eq. (5), where σ represents the sigmoid function and generates the output value between 0 and 1. The input gate (Γ_u) defines the extent (or degree) to which the current information is transformed into the memory cell. The forget gate (Γ_f) determines the degree to inform the cell state to discard the information or to forget the existing memory. Through the forgetting part of the current memory and adding new memory ($\tilde{c}^{<t>}$), the memory ($c^{<t>}$) is modified (restored). The gate (Γ_o) represents the output degree.

IV. RESULTS AND EVALUATION

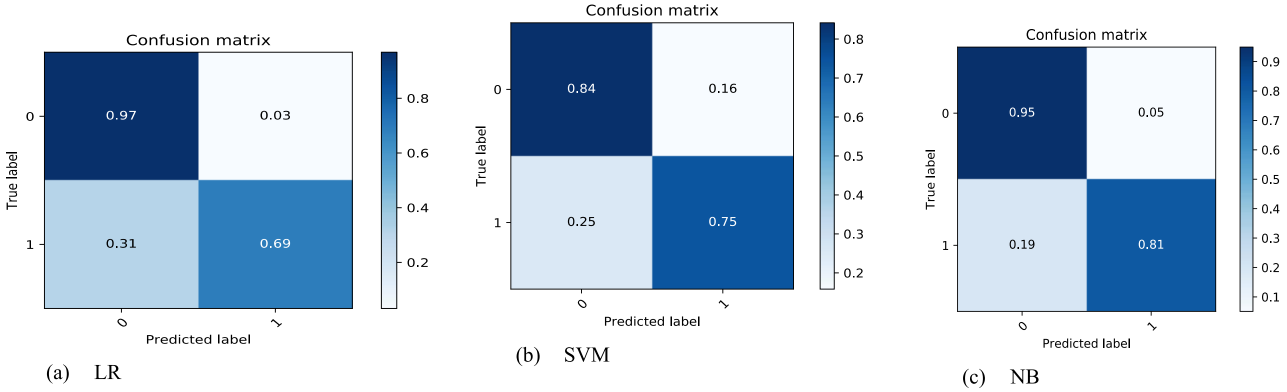
This section evaluates the effectiveness and potential of the proposed work by classifying Dengue disease. Also, we present the result achieved by performing various experiments in order to confirm all the research objectives. Experiments and results are conducted by using Python (3.6 version) and Anaconda framework.[3]

We trained our model using machine learning and deep learning classifiers techniques, discussed in Section 3.5, and evaluated each classifier performance on test data using the following performance evaluation measurements [67]: a) Accuracy, b) Precession, c) Recall, d) F1-Score, and e) Receiver Operating Character Curve (ROC Curve). The total average of the performance measures for each classifier using TF-IDF feature extraction is presented in Table 3. Where accuracy is the most commonly used classifier measurement of the performance metric. It is the measurement of all the correctly predicted actual classes such as actual positive and negative, referred to as Dengue infected or not infected in our case, in accordance to the total predictions. Moreover, precession determines out positive classes

[3] https://www.anaconda.com/distribution/

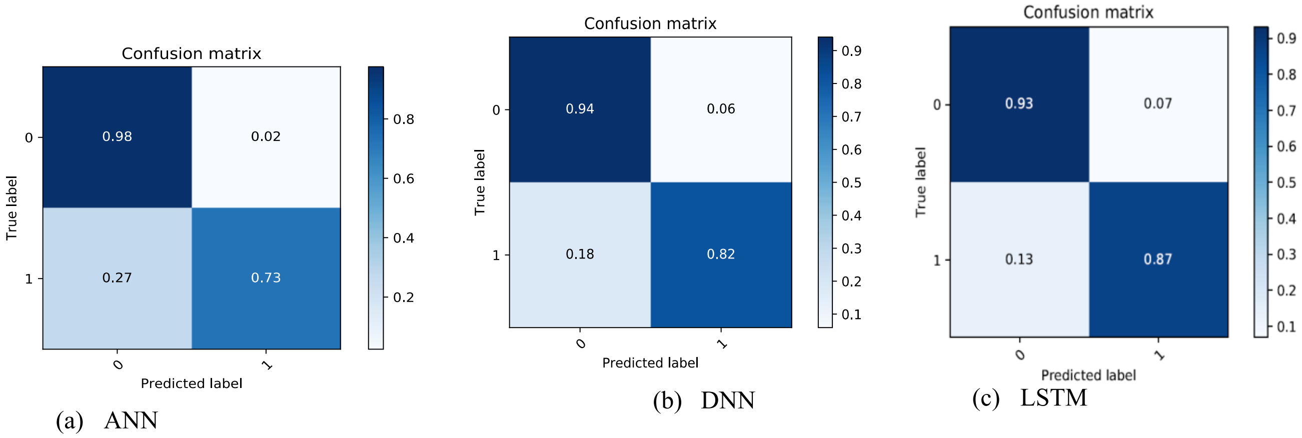
TABLE 3. Comparison of the proposed method with baseline methods using performance evaluation measurements.

Method	Train Accuracy (%)	Accuracy (%)	Precession (%)	Recall (%)	F1-Score (%)
LR	86.72	87.07	87.66	87.07	86.54
SVM	86.13	80.81	81.02	80.81	80.89
NB	94.60	89.95	89.93	89.95	89.82
ANN	92.30	88.92	89.48	88.92	88.53
DNN	93.76	89.82	89.75	89.82	89.72
LSTM	97.61	92.88	91.75	91.88	91.75



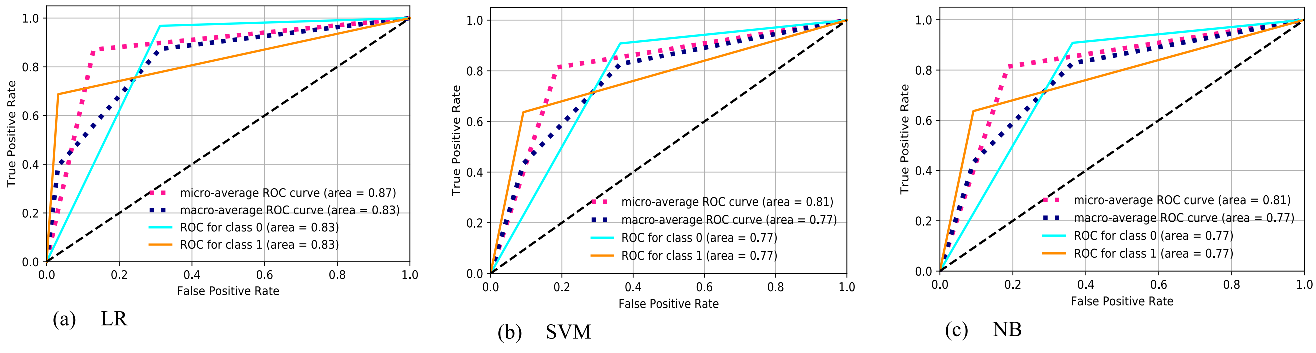
(a) LR (b) SVM (c) NB

FIGURE 3. Confusion matrix for (a) LR; (b) SVM; and (c) NB.



(a) ANN (b) DNN (c) LSTM

FIGURE 4. Confusion matrix for (a), ANN, (b) DNN, and (c) LSTM.

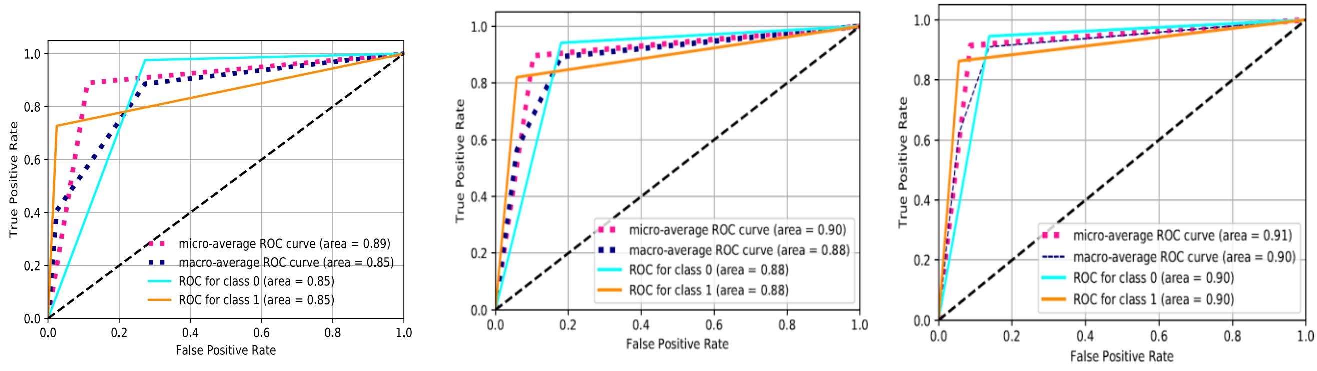


(a) LR (b) SVM (c) NB

FIGURE 5. ROC curve for (a) LR, (b) SVM, and (c) NB.

(i.e., Dengue infected people), which are predicted correctly, to analyze how many of those are positive (i.e., actually infected with Dengue). Recall presents the correctly predicted actual positive (Dengue infected people) is measured to all positive predictions. F1- Score also known as F measure and F score. It is a measure of test accuracy in supervised learning (binary classification) problem. F1 score is a weighted average function of precession and recall. The mean of the recall and precision is calculated via the F1 score. Following the results (see Table 3), it can be concluded that the LSTM classifier using TF-IDF features extraction technique achieved better performance results than the other classifiers.

We also graphically visualized the performance of the model by creating a confusion matrix (Figure 3 and 4) and ROC curve (Figure 5 and 6). A confusion matrix is a good solution to show results in two or more class classification problems as it further illuminates the classifiers' performance on test data and tries to compare the classified data according to their actual class label. However, ROC Curve is utilized to compute or measure of a binary classification model.



(a) ANN

(b) DNN

(c) LSTM

FIGURE 6. ROC curve for (a) ANN, (b) DNN, and (c) LSTM.

It illustrates that the higher the accuracy the better the model would be to detect the actual positive class as positive (Dengue infected (0) and ROC class (1) of the ROC curve for each classifier of machine learning and deep learning on TF-IDF features extraction technique given in Figure. 5 and Figure. 6, respectively.

From the above confusion matrices, we notice that the LSTM model (Figure. 4. C) has performed a better compared to other classifiers (Figure. 3 and Figure. 4), by predicting negative class labels when the tweet appeared negative (i.e., not infected people) but suffered from identifying positive class labels. This is due to the imbalance data as our training consisted of a large negative class. Consequently, the model learned from this class imbalance to provide a higher probability to a negative class label. Similarly, LSTM has also achieved higher accuracy for ROC (micro-average, macro-average, ROC class (0) and ROC class (1) compared to other classifiers.

Following the results above of the confusion matrix and ROC curve plots, it can be concluded that the LSTM classifier using TF-IDF features extraction technique achieved better performance results than other classifiers.

V. CONCLUSION

Social medical platforms are commonly used by people to express their opinions and disseminate personal information. People commonly share information related to health, science, business, art, daily life activities, etc. Using machine and deep learning techniques, we can extract useful information from this data and performs different analysis which can help to improve the quality of our life. In this paper, a model is proposed that extracts information from people's tweets. The tweet may be information about a person, whether he/she is actually infected with the Dengue disease or general information about the disease. The model uses RNN with a word embedding technique (TF-IDF). The experimental results demonstrate that the proposed model outperforms the current state-of-the-art machine and deep learning algorithms.

In future, we aim to evaluate the performances of the advanced word embedding techniques such as Word2Vec, Glove, and Fasttext in order to learn the semantic relationship among different words in social media text for better analysis.

REFERENCES

[1] M. A. Al-garadi, M. S. Khan, K. D. Varathan, G. Mujtaba, and A. M. Al-Kabsi, "Using online social networks to track a pandemic: A systematic review," *J. Biomed. Informat.*, vol. 62, pp. 1–11, Aug. 2016, doi: 10.1016/j.jbi.2016.05.005.

[2] D. Doran, K. Severin, S. Gokhale, and A. Dagnino, "Social media enabled human sensing for smart cities," *AI Commun.*, vol. 29, no. 1, pp. 57–75, Aug. 2015, doi: 10.3233/AIC-150683.

[3] S. C. Guntuku, D. B. Yaden, M. L. Kern, L. H. Ungar, and J. C. Eichstaedt, "Detecting depression and mental illness on social media: An integrative review," *Current Opinion Behav. Sci.*, vol. 18, pp. 43–49, Dec. 2017, doi: 10.1016/j.cobeha.2017.07.005.

[4] E. Cambria, "Affective computing and sentiment analysis," *IEEE Intell. Syst.*, vol. 31, no. 2, pp. 102–107, Mar. 2016, doi: 10.1109/MIS.2016.31.

[5] T. S. Raghavendra and K. G. Mohan, "Web mining and minimization framework design on sentimental analysis for social tweets using machine learning," *Procedia Comput. Sci.*, vol. 152, no. 1, pp. 230–235, 2019, doi: 10.1016/j.procs.2019.05.047.

[6] A. Hernandez-Suarez, G. Sanchez-Perez, K. Toscano-Medina, H. Perez-Meana, J. Portillo-Portillo, V. Sanchez, and L. García Villalba, "Using Twitter data to monitor natural disaster social dynamics: A recurrent neural network approach with word embeddings and kernel density estimation," *Sensors*, vol. 19, no. 7, p. 1746, Apr. 2019, doi: 10.3390/s19071746.

[7] P. S. Earle, D. C. Bowden, and M. Guy, "Twitter earthquake detection: Earthquake monitoring in a social world," *Ann. Geophys.*, vol. 54, no. 6, pp. 708–715, 2011, doi: 10.4401/ag-5364.

[8] T. D. Baruah, "Effectiveness of social media as a tool of communication and its potential for technology enabled connections: A micro-level study," *Int. J. Sci. Res. Publication*, vol. 2, no. 5, pp. 1–10, 2012. [Online]. Available: https://www.ijsrp.org

[9] K. R. Subramanian, "Influence of social media in personal communication," *ACADEMICA, Int. Multidiscip. Res. J.*, vol. 7, no. 9, p. 114, 2017, doi: 10.5958/2249-7137.2017.00093.3.

[10] K. Filo, D. Lock, and A. Karg, "Sport and social media research: A review," *Sport Manage. Rev.*, vol. 18, no. 2, pp. 166–181, May 2015, doi: 10.1016/j.smr.2014.11.001.

[11] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *J. Comput. Sci.*, vol. 2, no. 1, pp. 1–8, Mar. 2011, doi: 10.1016/j.jocs.2010.12.007.

[12] K. Jaidka, S. Ahmed, M. Skoric, and M. Hilbert, "Predicting elections from social media: A three-country, three-method comparative study," *Asian J. Commun.*, vol. 29, no. 3, pp. 252–273, May 2019, doi: 10.1080/01292986.2018.1453849.

[13] P. Salunkhe, A. Surnar, and S. Sonawane, ''A review: Prediction of election using Twitter sentiment analysis,'' *Int. J. Adv. Res. Comput. Eng. Technol.*, vol. 6, no. 5, pp. 723–725, 2017.

[14] E. Cambria, A. Hussain, T. Durrani, C. Havasi, C. Eckl, and J. Munro, ''Sentic computing for patient centered applications,'' in *Proc. IEEE 10th Int. Conf. Signal Process.*, Oct. 2010, pp. 1279–1282, doi: 10.1109/ICOSP.2010.5657072.

[15] A. J. Yepes, A. MacKinlay, and B. Han, ''Investigating public health surveillance using Twitter,'' in *Proc. Biomed. Nat. Lang. Process.*, 2015, pp. 164–170, doi: 10.18653/v1/w15-3821.

[16] S. Wakamiya, Y. Kawai, and E. Aramaki, ''Twitter–based influenza detection after flu peak via tweets with indirect information: Text mining study,'' *J. Med. Internet Res.*, vol. 20, no. 9, pp. 1–27, 2018, doi: 10.2196/publichealth.8627.

[17] A. Charalambous, ''Social media and health policy,'' *Asia-Pacific J. Oncol. Nurs.*, vol. 6, no. 1, pp. 24–27, 2019, doi: 10.4103/apjon.apjon_60_18.

[18] M. J. Paul and M. Dredze, ''Social monitoring for public health,'' *Synth. Lectures Inf. Concepts, Retr., Services*, vol. 9, no. 5, pp. 1–183, Aug. 2017, doi: 10.2200/s00791ed1v01y201707icr060.

[19] S. A. Moorhead, D. E. Hazlett, L. Harrison, J. K. Carroll, A. Irwin, and C. Hoving, ''A new dimension of health care: Systematic review of the uses, benefits, and limitations of social media for health communication,'' *J. Med. Internet Res.*, vol. 15, no. 4, pp. 1–16, 2013, doi: 10.2196/jmir.1933.

[20] H. Iso, S. Wakamiya, and E. Aramaki, ''Forecasting word model: Twitter-based influenza surveillance and prediction,'' M.S. thesis, Nara Inst. Sci. Technol., Ikoma, Japan, 2017.

[21] T. F. M. de Lima, R. M. Lana, T. G. De Senna Carneiro, C. T. Codeço, G. S. Machado, L. S. Ferreira, L. C. De Castro Medeiros, and C. A. Davis, ''DengueME: A tool for the modeling and simulation of dengue spatiotemporal dynamics,'' *Int. J. Environ. Res. Public Health*, vol. 13, no. 9, pp. 1–21, 2016, doi: 10.3390/ijerph13090920.

[22] P. Missier, A. Romanovsky, T. Miu, A. Pal, M. Daniilakis, A. Garcia, D. Cedrim, and L. da Silva Sousa, ''Tracking dengue epidemics using Twitter content classification and topic modelling,'' in *Current Trends in Web Engineering* (Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence) vol. 9881. 2016, pp. 80–92, 2016, doi: 10.1007/978-3-319-46963-8_7.

[23] A. Alessa and M. Faezipour, ''A review of influenza detection and prediction through social networking sites,'' *Theor. Biol. Med. Model.*, vol. 15, no. 1, pp. 1–27, Dec. 2018, doi: 10.1186/s12976-017-0074-5.

[24] C. de A. Marques-Toledo, C. M. Degener, L. Vinhal, G. Coelho, W. Meira, C. T. Codeço, and M. M. Teixeira, ''Dengue prediction by the Web: Tweets are a useful tool for estimating and forecasting Dengue at country and city level,'' *PLoS Neglected Tropical Disease*, vol. 11, no. 7, pp. 1–20, 2017, doi: 10.1371/journal.pntd.0005729.

[25] O. Titus Muurlink, P. Stephenson, M. Z. Islam, and A. W. Taylor-Robinson, ''Long-term predictors of dengue outbreaks in Bangladesh: A data mining approach,'' *Infectious Disease Model.*, vol. 3, no. 1, pp. 322–330, 2018, doi: 10.1016/j.idm.2018.11.004.

[26] M. J. Paul, A. Sarker, J. S. Brownstein, A. Nikfarjam, M. Scotch, K. L. Smith, and G. Gonzalez, ''Social media mining for public health monitoring and surveillance,'' in *Proc. Biocomputing*, Jan. 2016, pp. 468–479, doi: 10.1142/9789814749411_0043.

[27] Z. C. Lipton, J. Berkowitz, and C. Elkan, ''A critical review of recurrent neural networks for sequence learning,'' 2015, *arXiv:1506.00019*. [Online]. Available: http://arxiv.org/abs/1506.00019

[28] C. P. Medina and M. R. R. Ramon, ''Using TF-IDF to determine word relevance in document queries,'' in *Proc. 1st Instructional Conf. Mach. Learn.*, Piscataway, NJ, USA, 2003, pp. 133–142, doi: 10.15804/tner.2015.42.4.03.

[29] J. Violos, K. Tserpes, I. Varlamis, and T. Varvarigou, ''Text classification using the N-gram graph representation model over high frequency data streams,'' *Frontiers Appl. Math. Statist.*, vol. 4, pp. 1–19, Sep. 2018, doi: 10.3389/fams.2018.00041.

[30] T. Mikolov, K. Chen, G. Corrado, and J. Dean, ''Efficient estimation of word representations in vector space,'' in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Scottsdale, AZ, USA, 2013, pp. 1–12. [Online]. Available: http://arxiv.org/abs/1301.3781

[31] J. Pennington, R. Socher, and C. Manning, ''Glove: Global vectors for word representation,'' in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Doha, Qatar, 2014, pp. 1532–1543, doi: 10.3115/v1/D14-1162.

[32] H. Xue, Y. Bai, H. Hu, and H. Liang, ''Regional level influenza study based on Twitter and machine learning method,'' *PLoS ONE*, vol. 14, no. 4, pp. 231–253, 2019, doi: 10.1371/journal.pone.0215600.

[33] S. Wakamiya, M. Morita, Y. Kano, T. Ohkuma, and E. Aramaki, ''Tweet classification toward Twitter-based disease surveillance: New data, methods, and evaluations,'' *J. Med. Internet Res.*, vol. 21, no. 2, Feb. 2019, Art. no. e12783, doi: 10.2196/12783.

[34] A. Alessa and M. Faezipour, ''Preliminary flu outbreak prediction using twitter posts classification and linear regression with historical centers for disease control and prevention reports?: Prediction framework study,'' *JMIR Public Heal. Surveill.*, vol. 5, no. 2, pp. 1–17, 2019, doi: 10.2196/12383.

[35] K. Lee, A. Agrawal, and A. Choudhary, ''Real-time disease surveillance using Twitter data: Demonstration on flu and cancer,'' in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, Chicago, IL, USA, 2013, pp. 1474–1477, doi: 10.1145/2487575.2487709.

[36] 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 Med. Int. Health*, vol. 20, no. 7, pp. 840–863, Jul. 2015, doi: 10.1111/tmi.12498.

[37] K. Espina and M. R. J. E. Estuar, ''Infodemiology for syndromic surveillance of dengue and typhoid fever in the philippines,'' *Procedia Comput. Sci.*, vol. 121, no. 1, pp. 554–561, 2017, doi: 10.1016/j.procs.2017.11.073.

[38] C.-K. Wang, O. Singh, Z.-L. Tang, and H.-J. Dai, ''Using a recurrent neural network model for classification of tweets conveyed influenza-related information,'' in *Proc. Int. Work. Digit. Dis. Detect. Soc. Media (DDDSM)*, 2017, pp. 33–38. [Online]. Available: https://nlp.stanford.edu/projects/glove/

[39] M. Z. A. Bhuiyan, G. Wang, and Z. Fan, ''Smart world systems, applications, and technologies,'' *J. Netw. Comput. Appl.*, vol. 156, Apr. 2020, Art. no. 102553, doi: 10.1016/j.jnca.2020.102553.

[40] D. Gibert, C. Mateu, and J. Planes, ''The rise of machine learning for detection and classification of malware: Research developments, trends and challenges,'' *J. Netw. Comput. Appl.*, vol. 153, Mar. 2020, Art. no. 102526, doi: 10.1016/j.jnca.2019.102526.

[41] S. Rovetta, G. Suchacka, and F. Masulli, ''Bot recognition in a Web store: An approach based on unsupervised learning,'' *J. Netw. Comput. Appl.*, vol. 157, May 2020, Art. no. 102577, doi: 10.1016/j.jnca.2020.102577.

[42] J. Khan, I. Khan, A. Ghaffar, and B. Khalid, ''Epidemiological trends and risk factors associated with dengue disease in Pakistan (1980–2014): A systematic literature search and analysis,'' *BMC Public Health*, vol. 18, no. 1, pp. 1–13, Dec. 2018.

[43] M. Z. Yousaf, A. Siddique, U. A. Ashfaq, and M. Ali, ''Scenario of dengue infection & its control in Pakistan: An up-date and way forward,'' *Asian Pacific J. Tropical Med.*, vol. 11, no. 1, pp. 15–23, 2018, doi: 10.4103/1995-7645.223529.

[44] P. Guo, T. Liu, Q. Zhang, L. Wang, and J. Xiao, ''Developing a dengue forecast model using machine learning: A case study in China,'' *PLoS Neglected Tropical Disease*, vol. 11, no. 10, pp. 1–22, 2017, doi: 10.1371/journal.pntd.0005973.

[45] P. Ganeshkumar, M. V. Murhekar, V. Poornima, V. Saravanakumar, D. John, S. M. Mehendale, K. Sukumaran, and A. Anandaselvasankar, ''Dengue infection in India: A systematic review and meta-analysis,'' *PLoS Neglected Tropical Disease*, vol. 12, no. 7, pp. 1–29, 2018.

[46] M. O. Lwin, S. Vijaykumar, S. Foo, O. Noel, N. Fernando, and G. Lim, ''Social media-based civic engagement solutions for dengue prevention in Sri Lanka: Results of receptivity assessment,'' *Health Educ. Res.*, vol. 31, no. 1, pp. 1–11, 2016, doi: 10.1093/her/cyv065.

[47] P. Mutsuddy, S. T. Jhora, A. K. M. Shamsuzzaman, S. M. G. Kaisar, M. N. A. Khan, and S. Dhiman, ''Dengue situation in Bangladesh: An epidemiological shift in terms of morbidity and mortality,'' *Can. J. Infectious Disease Med. Microbiol.*, vol. 2019, no. 1, pp. 2017–2022, 2019, doi: 10.1155/2019/3516284.

[48] K. A. Agrupis, M. Ylade, J. Aldaba, A. L. Lopez, and J. Deen, ''Trends in dengue research in the Philippines: A systematic review,'' *PLoS Neglected Tropical Disease*, vol. 13, no. 4, pp. 1–18, 2019, doi: 10.1371/journal.pntd.0007280.

[49] J. A. Ayukekbong, O. G. Oyero, S. E. Nnukwu, H. N. Mesumbe, and C. N. Fobisong, ''Value of routine dengue diagnosis in endemic countries,'' *World J. Virol.*, vol. 6, no. 1, pp. 9–16, 2017, doi: 10.5501/wjv.v6.i1.9.

[50] N. Alsaedi, P. Burnap, and O. Rana, ''Can we predict a riot? Disruptive event detection using Twitter,'' *ACM Trans. Internet Technol.*, vol. 17, no. 2, pp. 1–26, May 2017, doi: 10.1145/2996183.

[51] H. Saif, M. Fernandez, Y. He, and H. Alani, ''Evaluation datasets for Twitter sentiment analysis a survey and a new dataset, the STS-Gold,'' in *Proc. CEUR Workshop*, vol. 1096, 2013, pp. 9–21.

[52] B. Amina and T. Azim, "SCANCPECLENS: A framework for automatic lexicon generation and sentiment analysis of micro blogging data on China Pakistan economic corridor," *IEEE Access*, vol. 7, pp. 133876–133887, 2019, doi: 10.1109/ACCESS.2019.2940528.

[53] R. Artstein, "Inter-annotator agreement," in *Handbook of Linguistic Annotation*. Dordrecht, The Netherlands: Springer, 3017, pp. 297–313. [Online]. Available: http://link.springer.com/10.1007/978-94-024-0881-2_11

[54] M. L. McHugh, "Interrater reliability: The kappa statistic," *Biochemia Medica*, vol. 22, no. 3, pp. 276–282, 2012. [Online]. Available: https://hrcak.srce.hr/89395

[55] J. L. Fleiss, B. Levin, and M. C. Paik, "The measurement of interrater agreement," in *Statistical Methods for Rates and Proportions*. Hoboken, NJ, USA: Wiley, 2004, pp. 598–626, doi: 10.1002/0471445428.ch18.

[56] B. Han and T. Baldwin, "Lexical normalisation of short text messages: Makn sens a #Twitter," in *Proc. 49th Annu. Meet. Assoc. Comput. Linguistics Hum. Lang. Technol.*, vol. 1, 2011, pp. 368–378.

[57] S. Bird, E. Klein, and L. Edward, *Natural Language Processing With Python: Analyzing Text With the Natural Language Toolkit*. Newton, MA, USA: O'Reilly Media, 2009.

[58] S. Nirmal and T. Verma, "E-Mail spam detection and classification using SVM and feature Extraction," *Int. J. Advance Res., Ideas Innov. Technol.*, vol. 3, no. 3, pp. 1491–1495, 2017.

[59] C.-Y.-J. Peng, K. L. Lee, and G. M. Ingersoll, "An introduction to logistic regression analysis and reporting," *J. Educ. Res.*, vol. 96, no. 1, pp. 3–14, Sep. 2002, doi: 10.1080/00220670209598786.

[60] C. Cortes and V. Vapnik, "Support-vector networks," *IEEE Expert. Syst. Appl.*, vol. 7, no. 5, pp. 63–72, Sep. 1995, doi: 10.1007/BF00994018.

[61] D. D. Lewis, "Naive (Bayes) at forty: The independence assumption in information retrieval," in *Proc. Eur. Conf. Mach. Learn.*, 1998, pp. 4–15.

[62] K. Saravanan and S. Sasithra, "Review on classification based on artificial neural networks," *Int. J. Ambient Syst. Appl.*, vol. 2, no. 4, pp. 11–18, Dec. 2014, doi: 10.5121/ijasa.2014.2402.

[63] V. A. Maksimenko, S. A. Kurkin, E. N. Pitsik, V. Y. Musatov, A. E. Runnova, T. Y. Efremova, A. E. Hramov, and A. N. Pisarchik, "Artificial neural network classification of motor-related EEG: An increase in classification accuracy by reducing signal complexity," *Complexity*, vol. 2018, pp. 1–10, Aug. 2018, doi: 10.1155/2018/9385947.

[64] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Netw.*, vol. 5, no. 2, pp. 157–166, Mar. 1994, doi: 10.1109/72.279181.

[65] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[66] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014, *arXiv:1412.3555*. [Online]. Available: http://arxiv.org/abs/1412.3555

[67] D. M. W. Powers, "Evaluation: From precision, recall and F-measure to Roc, informedness, markedness & correlation," *J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 37–63, 2011.

SAMINA AMIN received the M.Sc. degree in computer science from the Institute of Computing, Kohat University of Science and Technology, Kohat, Pakistan, where she is currently pursuing the M.S. degree. Her research interests include machine learning, deep learning, and social media analysis. She received the Gold Medal from the Kohat University of Science and Technology.

M. IRFAN UDDIN (Member, IEEE) received the B.Sc. and M.Sc. degrees in computer science from the University of Peshawar, and the M.S. degree leading to the Ph.D. degree with the University of Amsterdam, The Netherlands. He has been involved with academia and research, since 2005. He was a Research Associate with the University of Amsterdam and the University of Turin, Italy. He was an Assistant Professor with Al Yamamah University, Saudi Arabia. He is currently an Assistant Professor with the Institute of Computing, Kohat University of Science and Technology, Kohat, Pakistan. His research interests include machine learning, data science, deep learning, convolutional neural networks, reinforcement learning, computer vision, and parallel programming. He serves as a Reviewer for different journals.

SAIMA HASSAN received the M.Sc. degree in computer science from the University of Peshawar, Pakistan, in 2003, and the M.Sc. and Ph.D. degrees in information technology from the Universiti Teknologi PETRONAS, Malaysia, in 2013 and 2016, respectively. She is currently an Assistant Professor with the Institute of Information Technology, Kohat University of Science and Technology, Kohat, Pakistan. Her research interests include time series forecasting, artificial neural networks, and application of computational intelligence techniques to load forecasting.

ATIF KHAN received the M.Sc. degree in computer science from the University of Peshawar, Pakistan, in 2004, and the Ph.D. degree in computer science (text mining) from the Universiti Teknologi Malaysia (UTM), Johor Bahru, Malaysia, in 2016. Since 2016, he has been an Assistant Professor with Islamia College Peshawar, Pakistan. His current research interests include data mining, text mining, sentiment analysis and opinion mining, recommender systems, and machine learning. He is a technical committee member in many international conferences. He was a recipient of the Best Student Award and the Pro-Chancellor Award from UTM, during the Ph.D. degree, for his excellent contribution in the field of text mining. He serves as an Associate Editor for *ACM Transactions on Asian and Low-Resource Language Information Processing*. He also serves as a Reviewer for many international conferences and journals.

NIDAL NASSER (Senior Member, IEEE) received the Ph.D. degree from the School of Computing, Queen's University, Canada, in 2004. He is currently a Professor of software engineering with the College of Engineering, Alfaisal University, Saudi Arabia, where he is also the Founder and the Director of the Internet of Things Research Laboratory. He has authored 180 journal publications, refereed conference publications, and book chapters, in the area of wireless communication networks and systems. He has also given keynote speeches and tutorials in major international conferences. He received the Best Paper Award from the different IEEE and the international conferences, such as the IEEE ICC, in 2014.

ABDULLAH ALHARBI received the Ph.D. degree from the University of Technology Sydney, Australia. He is currently an Assistant Professor with the Information Technology Department, Taif University. His research interests include human–computer interaction, information systems, cybersecurity, and data science.

HASHEM ALYAMI received the bachelor's degree in computer Science from Taif University, Saudi Arabia, in 2007, the master's degree in secure computer system from the University of Hertfordshire, U.K., and the Ph.D. degree from the University of Reading, U.K. He is currently an Assistant Professor with the Computer and Information Technology College, Taif University. His research interests include cybersecurity, artificial intelligent, and data science.