2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT) | 978-1-6654-1924-6/20/$31.00 ©2020 IEEE | DOI: 10.1109/WIIAT50758.2020.00131

A Social Media Time-Series Data Analytics Approach for Digital Epidemiology

Md. Aslam Parwez
Department of Computer Science
Jamia Millia Islamia
New Delhi, India
aslamparwez.jmi@gmail.com

Muhammad Abulaish, *SMIEEE*
Department of Computer Science
South Asian University
New Delhi, India
abulaish@sau.ac.in

Jahiruddin
Department of Computer Science
Jamia Millia Islamia
New Delhi, India
jahiruddin@jmi.ac.in

***Abstract*—Various events and their perspectives around the world are discussed or posted at every moment on social media platforms like `Twitter` in near real-time, forming an enriched repository of information as historical records or time series. These include people's sentiments, emotions, opinions, and other information such as situational aspects of the spreading of a particular disease, ailment, or a population explosion of some vectors or pathogens. Exploring and harnessing such information about a disease for surveillance to prevent and control its spreading or becoming epidemic or pandemic is worthwhile for a country or the world. In this paper, we correlate tweeting activity with the reported disease cases, and take advantage of the predictive power of neural networks and auto-regressive models to estimate disease incidences for the current week (*aka nowcasting*) considering the social media data and the disease case counts reported by the Government agencies. We propose Long Short-Term Memory (LSTM) network models and autoregressive moving average models with two channels of inputs to incorporate social media and historic disease case count data for predicting current disease case counts. We employ various LSTM network models and autoregressive moving average models to estimate the current week's disease case count and compared their performance considering tweets as exogenous input to these models. The experimental results establish the efficacy of the LSTM network models with dynamically merged inputs for predicting disease case count with least prediction error.**

***Index Terms*—Digital epidemiology, Disease surveillance, Time series forecasting, Social media, Neural network, LSTM, Nowcasting.**

I. Introduction

In recent decades, the world has changed expeditiously with exceptional growth in population together with the menace of emerging and re-emerging infectious diseases. As the infectious diseases spread rapidly throughout the population, their early warning signs emerge in different media in near real-time in terms of news, blogs, microblogs, bulletin boards, and other social media posts. Collecting and analyzing data from these media sources could be helpful for surveillance systems to provide information on early warning about the diseases so that suitable measures could be taken to prevent any adverse situation to become an epidemic or pandemic [1]. Traditionally many surveillance techniques and systems have been developed and deployed. Most common among them include *vital statistics* about birth and death [2], registries for particular conditions or defects, a routine survey of the population, disease reporting to state and national agencies, adverse event (e.g., drugs and vaccines) surveillance [3], sentinel surveillance to report disease condition or cases (e.g., influenza-like-illness in the United States) [2], zoonotic disease surveillance to detect infected animals [4], laboratory test data, and syndromic surveillance [5] using clinical signs and symptoms about a disease.

Nowadays, the growth and easy access to the internet has increased its penetration in various regions of the world, and social media has become ubiquitous in our society. Today social media is an essential medium of social interaction through mobile and web-based interfaces. Social media sites, such as `Twitter` provides a distinctive platform where people disclose and share their personal health information. Such information is distributed in near real-time and accessible to communities irrespective of location and time. Social media contents are of shorter length (e.g., tweet-length limited to 280 characters, earlier it was 140 characters) and dynamic in nature that may represent an opinion, sentiment, trending event, topic, discussion, and health information. Monitoring and analyzing these contents would help to understand user behavior, what is trending topic, what circumstances peoples are concerned with, and which disease is spreading in a particular location.

Many researchers admit that tracking social media data for influenza outbreak detection is more efficacious and authentic than the traditional method of reporting based on sentinel surveillance [6]–[9]. Despite the increasing interest of researchers and academicians to exploit social media data and harness its potential for nowcasting and forecasting disease outbreak detection, public health professionals are unwilling to leverage such insights. Traditional surveillance systems depend on public health authorities involved in hospitals, laboratories, or over-the-counter drug sales to collect and report disease incidence data at local, regional, and national levels, which cause undesirable delays and hinders timely reporting of disease cases. As a result, an average of two weeks delay happens in the occurrence of diseases and their official notification [10]. In countries like India, where health is a state subject with many states not reporting timely to the nodal agencies, such delay is a major issue [11].

This paper takes advantage of the predictive power of neural networks and auto-regressive models for estimation of disease

978-1-6654-1924-6/20/$31.00 ©2020 IEEE
DOI 10.1109/WIIAT50758.2020.00131

case count based on social media data and the historic disease incidence data reported by the Govt. agencies. It provides an important surveillance application called *nowcasting*, which is a form of forecasting in which we predict the present level of the disease case counts that is yet unknown. We employ four neural network-based Long Short-Term Memory (LSTM) network models and two autoregressive moving average models for *nowcasting* of disease incidences and compared their performance. We observe that the models learned on social media data together with disease case count data outperform other models. Moreover, the LSTM models with combined social media and historic case count data perform considerably better than autoregressive moving average models. Thus, we anticipate that LSTM models with inputs from social media data and historic case count data can play a pivotal role in monitoring and detecting possible threats to public health and capturing early signals of any epidemic or pandemic. In addition, our approach is generic and applicable to any infectious disease for a similar purpose of forecasting. For experimental evaluation, we have considered dengue as a case study to monitor and predict the level of dengue disease incidences in India.

The remaining part of the paper is organized as follows. Section II presents a review of the existing state-of-the-arts on disease surveillance. Section III presents a preliminary study on various auto regressive models, deep learning-based LSTM networks, and other background details. Section IV explains the proposed approach for nowcasting disease incidences. The experimental setup and performance evaluation are discussed in section V. Finally, section VI concludes the paper with future directions of research.

II. Related Work

Monitoring the spreading of diseases to minimize the potential damage caused by their outbreak has been in focus for many years [2], [3], [12]. Emphasis is given on two types of surveillance, of which a traditional approach based on case-based reporting is termed as *passive surveillance* while a recent approach based on Internet-based surveillance is termed as *active surveillance* [12]. Internet/Web-based surveillance is based on news articles, RSS feeds, search engine queries, and social media content. Social media nowadays has become a center of attraction for the research community because of its live streaming, diverse content, real-time availability of enormous data. Many studies have been accomplished to analyze sentiments, identify political trends, measure the intensity of disease outbreaks and public health behaviours [13]–[16]. Social media in general and `Twitter`, in particular, have attracted many researchers for studying public health perceptions.

In [13], authors reported the link between Google trends search metrics (i.e., the terms correlated to the disease along with their frequencies) and Australian weekly notification data about some diseases. They found a strong correlation between search metrics and disease notification data. They applied the concept of a linear model to calculate query lag between search metrics and disease notification to have the best model to predict one week or two weeks of disease incidences. Thapen et al. [14] proposed DEFENDER system by integrating social media and news media data for outbreak detection, situational awareness, nowcasting to count current level of disease activity, and forecasting to predict future symptom counts based on observation of symptomatic people movement from neighbouring regions to an area. In [17], authors used Wikipedia access logs and disease incidence data and language as a proxy for the location to produce an effective disease monitoring and forecasting system.

The majority of works in disease surveillance using social media have emphasized on infectious diseases of which influenza-like illness [6], [7], [18], [19] has received the greatest attention. Another notable disease received researchers' attention for surveillance is the dengue fever [20]–[22]. Researchers have also examined other diseases using social and news media data that include cholera [23], malaria [24], ebola virus outbreak [25], [26], zika virus outbreak [27], [28], and others. Non-infectious diseases have also received some attention of researcher of which cancer [29], [30], asthma [31], diabetes [32], human immunodeficiency virus [33], [34], to name a few.

Using social media data for surveillance, however, is a subject of major concerns as limitations have been pointed out by researchers [35], [36], especially, after the failure of Google Flu Trend [37]. Social media data are posted or reported by individuals, which may misreport or under-report their health issues resulting in over or underestimation that may aggravate the surveillance [38], [39]. There are other factors like biases towards representativeness of the data that include the user's demographics and lack of domain expertise, bot accounts that may influence people's perception towards a disease or drug, and privacy concerns that may affect the reliability of surveillance for decision making and health interventions. Despite these limitations, numerous studies have continually revealed the strength of social media data in public health surveillance, more particularly for disease surveillance. Social media monitoring cannot replace traditional disease surveillance. However, it can act as a complementary tool for traditional systems by addressing the knowledge gaps by reducing the cost and latency of collecting information and their analysis.

Surveillance systems based on social media data are intended to sense any abnormalities in the volume of user-generated content associated with health-related adverse events. The abnormalities can be detected using statistical models such as time series models, which are fitted on time series data to forecast the disease epidemics condition. Time series forecasting in domains such infectious diseases are traditionally done using Auto Regressive Moving Average models [40], [41]. Recently, deep learning based models have gained popularity for text information processing [42], including time series forecasting in different domains. Researchers in [43], [44] have shown that the deep learning based models such as LSTM outperform auto regressive algorithms for time series forecasting tasks.

III. Preliminaries

Time series data represents a set of data points at a consecutive time arranged in chronological sequence and common in many fields, including science, engineering, business, and economics. We need to fit an appropriate model to the time series data and estimate parameters such that the learned mathematical model can be used to forecast or predict the future values. Based on the predicted value, the policymakers take precautionary measures and make interim decisions. Therefore, selecting an appropriate model is essential to represent the time series data so we could use the fitted model for forecasting.

There are two important linear time series models, namely *Auto Regressive* (AR) [45] and *Moving Average* (MA) [45] models. The AR model regresses a time series value from the past values (lagged values) of the given time series. It can be represented using equation 1, where the dependent variables y_{t-1}, y_{t-2}, ..., y_{t-k} are the response variables of the past k time periods, ϵ_t is the error term (white noise) in regression model, β_0 is a constant term, and β_1, β_2, β_3, ..., β_k are coefficients of the past k response variables, and k is the order of AR model.

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \beta_3 y_{t-3} + ... + \beta_k y_{t-k} + \epsilon_t \quad (1)$$

The MA model, on the other hand, uses past forecast errors to make current forecast using a regression like model represented by equation 2, where μ is the mean of the series, ϵ_{t-1}, ϵ_{t-2}, ϵ_{t-3}, ..., ϵ_{t-k} represent past error terms, ϵ_t is the current error term, and θ_1, θ_2, θ_3,...,θ_k are coefficients of the past k error terms.

$$y_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \epsilon_{t-3} + ... + \theta_k \epsilon_{t-k}. \quad (2)$$

The combination of these two models results in more sophisticated models that include both the auto regressive and moving average components, which are discussed in the following subsections.

A. ARMA, ARIMA, and ARMAX Methods

Auto Regressive Moving Average (ARMA) model: It involves two parts, namely Auto Regressive (AR) and Moving Average (MA) parts of which the AR part regresses the past values to time series while the MA part models the prior error terms. The ARMA model is represented by equation 3, where p and q represent the order of AR and MA parts, respectively.

$$y_t = \beta_0 + \sum_{i}^{p} \beta_i y_{t-i} + \epsilon_t + \sum_{j}^{q} \theta_j \epsilon_{t-j} \quad (3)$$

The ARMA model can be fitted on stationary times series. A series is said to be stationary if its mean is constant and does not vary as a function of time. Similarly, the variance of the stationary series does not depend on time. If the series does not meet the stationary criteria, we need to make it stationary using suitable techniques that include *detrending*, *differencing*, and *seasonality*. The most common technique for making a series stationary is *differencing*, in which we model the differences of terms instead of actual terms. The *differencing* is referred to as integration part (I) in the ARMA model, and this gives rise to a new model called Auto Regressive Integrated Moving Average model which is discussed as follows:

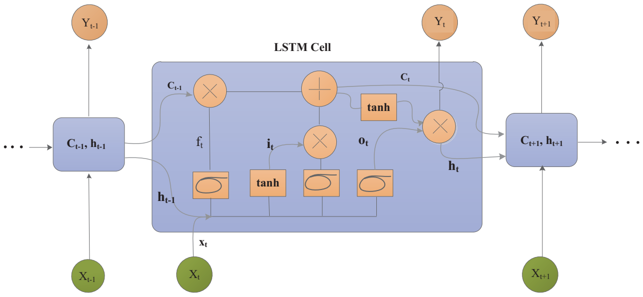

Fig. 1: A Long Short-Term Memory network with LSTM units

Auto Regressive Integrated Moving Average (ARIMA) model: It is a generalized version of the ARMA model with an additional component called integration (I). It is denoted by ARIMA(p,d,q), where the parameters p represents lag order, d represents the degree of differencing, and q represent the moving average window. One needs to prepare the data according to the differencing order to construct the ARIMA model.

Auto Regressive Moving Average with Exogenous Inputs (ARMAX) model: ARMAX is a generalized version of the ARMA model that incorporates exogenous input variable X. This model contains p auto regressive terms, q moving average terms, and a linear combination of b terms of an external time series. The model can be represented by equation 4, where d_t represents the external time series and the $\eta_1, \eta_2, ..., \eta_k$ represent parameters of the series d_t.

$$y_t = \beta_0 + \sum_{i}^{p} \beta_i y_{t-i} + \epsilon_t + \sum_{j}^{q} \theta_j \epsilon_{t-j} + \sum_{k}^{b} \eta_k d_{t-k} \quad (4)$$

B. Long Short-Term Memory Network

Long Short-Term Memory (LSTM) neural network is a recurrent neural network that has embedded LSTM units with each unit comprising a memory cell to store information and three gates including an input, forget, and output gate to regulate information flow within the memory cell [46]. The input gate provides new information as input to the cell and regulates the recent information using activation function, while the forget gate discards some information from the existing content of the memory cell. Likewise, the output gate determines the amount of information to be transmitted to the next hidden state. LSTM networks store recent short-term history as activation of neurons and long-term history as weights, which get modified during backpropagation. The memory cell learns through the iterative process of guessing, back-propagating errors, and modifying weights through gradient descent for regulating information, preventing any vanishing and explosion of gradients. The values of input, output, forget gates, and the candidate memory cell at a time-step t with input vector w_t are updated using equations 5, 6, 7, 8, 9, 10. In these equations, $\odot$ represents element-wise

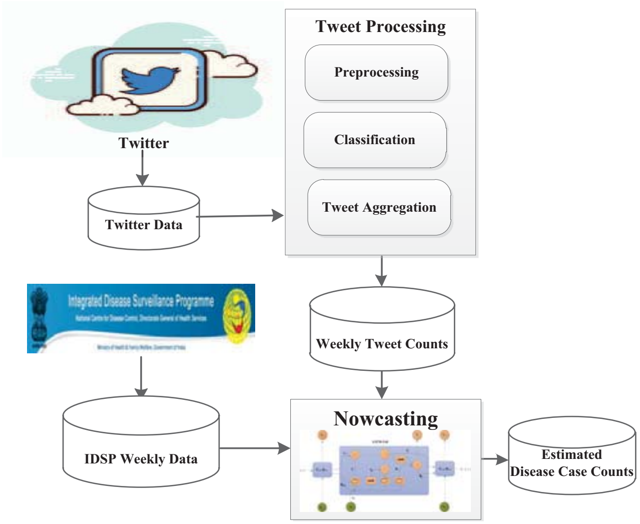

Fig. 2: Work-flow of the proposed disease surveillance approach

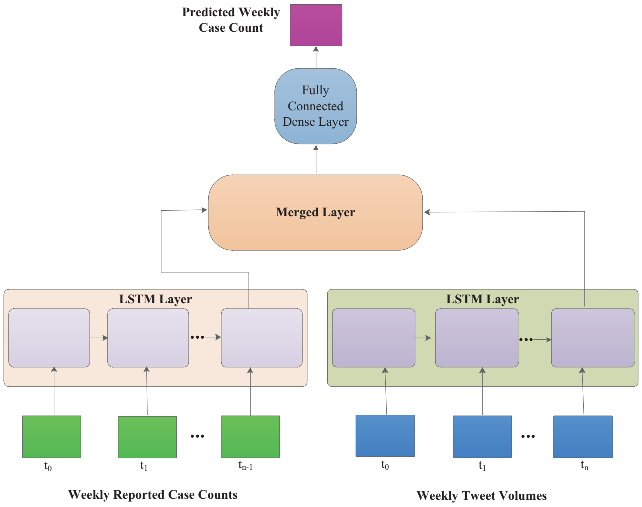
Predicted Weekly Case Count
Fully Connected Dense Layer
Merged Layer
LSTM Layer
LSTM Layer
t0
t1
tn-1
t0
t1
tn
Weekly Reported Case Counts
Weekly Tweet Volumes

Fig. 3: LSTM network with two channels of inputs for nowcasting of disease cases. The merging of outputs from the LSTM layers is done either by concatenation or by average resulting in two different models `TweetX-LSTMConcat` and `TweetX-LSTMAverage`, respectively.

multiplication, σ represents sigmoid function and W_i, b_i, W_f, b_f, W_o, b_o represent input, forget and output gates parameters. The final hidden vector representing high level features is fed into dense layer.

$$i_t = \sigma(W_i.[h_{t-1}; w_t] + b_i) \tag{5}$$

$$f_t = \sigma(W_f.[h_{t-1}; w_t] + b_f) \tag{6}$$

$$o_t = \sigma(W_o.[h_{t-1}; w_t] + b_o) \tag{7}$$

$$g_t = tanh(W_r.[h_{t-1}; w_t] + b_r) \tag{8}$$

$$c_t = i_t \odot g_t + f_t \odot c_{t-1} \tag{9}$$

$$h_t = o_t \odot tanh(c_t) \tag{10}$$

LSTMs are very effective in modeling sequential data because of their capability of learning and remembering long input sequences, and hence, they have been successfully applied for sequence labeling, speech recognition, image captioning, and language modeling [47].

IV. Proposed Nowcasting Approach

In this section, we present an approach of nowcasting disease incidences based on social media data and the actual disease case count data reported to government health agencies. Figure 2 presents an overall work-flow of the proposed nowcasting approach. It involves data collection, processing, and use of time series models for nowcasting disease cases using `Twitter` data and the data reported through Integrated Disease Surveillance Programs (IDSP) by the Govt. of India.

A. Dataset Acquisition

We have used two relevant data sources for event detection and nowcasting of the number of disease incidences. As a case study, we have considered dengue, which is endemic in India and most prevalent in tropic and sub-tropic regions because of favourable climatic conditions for the breeding of vectors and pathogens in these regions. The following subsections present the details of these data sources.

1) Twitter Data: We collected tweets originated from India during June 18, 2017, to July 14, 2018, using the `Twitter4J` Java library. During the given period, we collected 30 million tweets using a geographic bounding box around India. However, for experimentation, we used only English language tweets associated with dengue disease that consists of 0.239 million tweets.

2) IDSP Case Count Data: Integrated Disease Surveillance Program[1] (IDSP) is a significant initiative by National Health Mission for monitoring and reporting epidemic-prone disease as a part of National Health Programme for all States and UTs of India. It uses trained Rapid Response Team (RRTs) based on the decentralized laboratory to monitor disease trends and to identify and report the disease outbreaks in its nascent stage. IDSP reports the weekly case count and deaths due to 22 disease/illness reported to it by states and UTs. We collected weekly dengue case count data from June 2017 to July 2018 for our experimentation. We represent the weeks from 17W23 to 18W28 to reflect the year and week number of that year such that the epidemiological week 17W23 represents the 23^{rd} week of the year 2017.

B. Tweet Pre-processing and Aggregation

To address the noise-related issues of the data, we pre-processed the raw tweets and transformed them into a readily processable form appropriate for representing as input to the

[1] https://idsp.nic.in/index.php

machine learning algorithm for classification. The tweet texts are pre-processed by eliminating URLs, user mentions, stop words, punctuation, numerals, and extra spaces, and the text tokens are converted to lowercase. We then employed a basic CNN-based classifier discussed in [48] to identify disease-related tweets. The disease-related tweets thus collected are aggregated into weekly tweet counts for the given disease and thus formulated a time series data for experimentation.

C. Outbreak Patterns Detection

To detect outbreak patterns, we examined the distribution or trends of tweets volumes generated per week, and their correlation with the dengue case counts throughout the entire data collection period discussed in section IV-A. Figure 4 presents the weekly dengue cases reported through IDSP, and the tweets count during the same epidemiological weeks (henceforth mentioned as Epi weeks). The graphs and the spikes in this figure suggest that there is a close association between the volume of tweets and the reported dengue cases. Further, it can be observed from figure 4 that the tweet volume and the disease case count shares a similar trend. In figure 4 the case count has a different numerical range with respect to the tweet volume per week, as marked on the y-axis. We can observe from figure 4 that the dengue case counts show peaks in June, July, and August during Epi weeks 17W27, 17W29, 17W31, and 17W34 for the year 2017. It further shows peaks during weeks 17W38 to 17W40. Thereafter, we see a sharp decrease in the number of dengue cases during November and December 2017, and it continues to be low until April 2018. However, from May 2018 onward, we again see an increase in the number of dengue cases and some peaks (18W22 and 18W25) during June 2018. The highest peak is seen in July 2017 in Epi Week 17W31, with a total number of dengue cases in that week exceeds 5,000 and the total number of dengue cases detected during July 2017 to be 7,921. Further, we can observe from figure 4 that the trends of tweet volumes are also able to detect most of these outbreaks by depicting surge in tweets during the peak periods. However, the tweet volumes were unable to capture certain major peaks, like the peak in Epi week 18W22 in May 2018. Further, we can also observe that spikes in tweet volumes per Epi Weeks occur immediately before the dengue case count reports, and there are one to two weeks lag between spikes in tweets and the peak value in immediate dengue case count reports.

D. Nowcasting

The purpose of nowcasting is to estimate the current case count using the past reported disease case count by Govt. agencies available till the previous time point T' and the tweet count up to the current time point T. As the reports published by Govt. agencies (e.g., IDSP) generally gets delayed by one week to one month. Considering delay of one time point say $\Delta = T' - T = 1$, we can assume that the current tweet count available till T will help in estimating the current case count. This can be considered as one step ahead estimation of the disease case count.

1) Auto Regressive Moving Average Models: We consider two auto regressive moving average models – i) `CaseCount-ARMA` and ii) `Tweet-ARMAX` models. The `CaseCount-ARMA` model is based on the ARMA model, and `Tweet-ARMAX` is based on the ARMAX model discussed in section III. The `CaseCount-ARMA` model takes only one input time series data, which is the historical case count data reported by the Govt. health agencies. On the other hand, the `Tweet-ARMAX` model incorporates two input time series data with external input as the weekly tweet volumes available until time point T.

2) Long Short-Term Memory Network Models: We consider four different LSTM models – i) a `CaseCount-LSTM`, ii) `TweetX-LSTMConcat`, iii) `TweetX-LSTMAverage`, and iv) `MultiVariate-LSTM`. The `CaseCount-LSTM` is a simple LSTM model that considers a LSTM layer followed by a dense layer to predict the disease case count. It considers only one input channel for nowcasting. `TweetX-LSTMConcat` and `TweetX-LSTMAverage` are merged forms of LSTM models that take two channels of inputs and the outputs of LSTM layers from two channels are dynamically merged by concatenation and average respectively before they are padded to the fully connected dense layer. The `TweetX-LSTMConcat` model takes two channels of inputs to two different LSTM layers with similar structure and parameters, of which one for previous disease case counts and the other for weekly tweet counts. The outputs of the LSTM layers are merged by concatenation before they are passed to a fully connected dense layer to predict the disease case counts. Similarly, the `TweetX-LSTMAverage` model takes two channels of inputs to two different LSTM layers with similar structure and parameters of which one for previous disease case counts and the other for weekly tweet counts. The outputs of the LSTM layers are merged by element-wise average before they are passed to a fully connected dense layer to predict the disease case counts. The `MultiVariate-LSTM` on the other hand takes two parallel series as input to a single LSTM model and predicts the output for the next time steps based on the two series. `MultiVariate-LSTM` is different from the merged models in the sense that it has a single input of two parallel series and have only one LSTM layer followed by a fully connected dense layer. We employed the dense layer in all these models without any activation as we predict the numerical value directly, just like a regression task.

V. Experimental Setup and Results

For experimental setup, we used python library `keras`[2] especially designed for deep learning models, and an important python packages `statsmodels`[3], which provides functions and classes helpful for estimating various statistical models, exploring statistical data, and conducting statistical tests. All the experiments are conducted using python 3.6.5. For evaluation, we divided the available case count data and the tweet volumes

[2] https://keras.io/api/

[3] https://www.statsmodels.org/stable/index.html

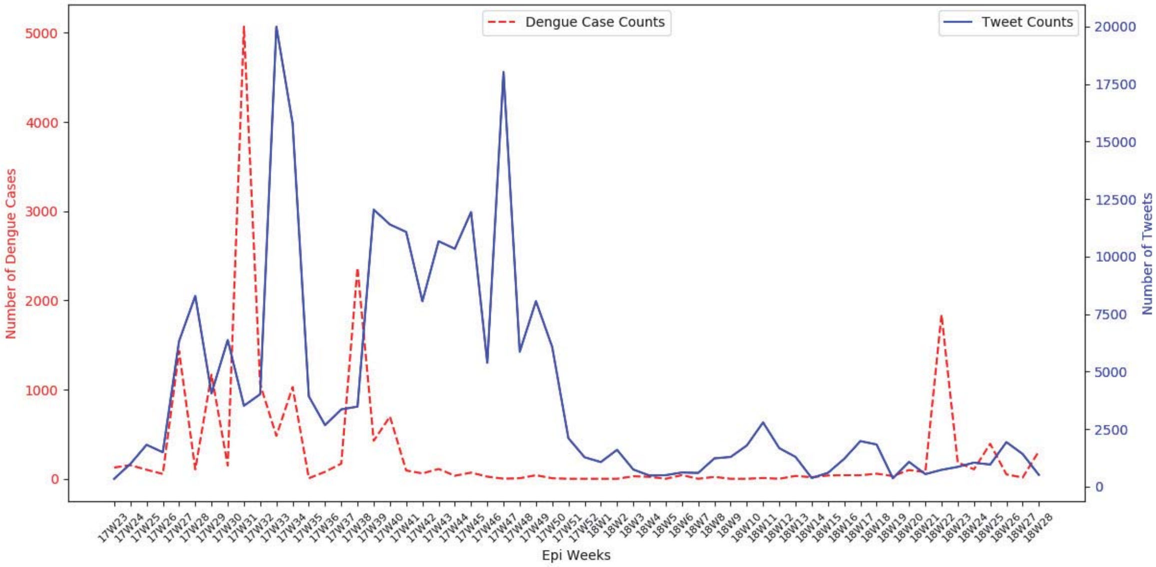

Fig. 4: Graph showing the relationship between the tweets count and the reported dengue cases. *Note*, the dengue case counts are at different numerical range with respect to the tweets count.

into training and validation sets (training : validation ratio as $80 : 20$) and used the validation set to evaluate the estimated case count. The time slot for the training period is from June 5, 2017, to April 22, 2018, and for evaluation is from April 23, 2018, to July 16, 2018. The performance of nowcasting is evaluated based on ground truth reports of the IDSP data.

For CaseCount-ARMA, and Tweet-ARMAX models, we used the orders p=1, and q=1 for these models. Similarly, for deep learning-based LSTM models – CaseCount-LSTM, TweetX-LSTMConcat, TweetX-LSTMAverage, and Multivariate-LSTM, we set the number of LSTM units to 256 neurons and used ReLU [49] activation function. The models are trained for 500 epochs using Adam [50] optimizer to update parameters. We applied the mean square error loss function to minimize the error and used the fully connected dense layer without any activation because we are predicting numerical value like regression tasks.

A. Nowcasting Evaluation

We compared the performance of the auto regressive moving average models and LSTM models for nowcasting task. In the case of CaseCount-ARMA model, we incorporated the ARMA model into the IDSP disease case count data to estimate the current case count during the validation period. CaseCount-ARMA model does not use any exogenous input for the estimation of case counts. Similarly, the CaseCount-LSTM model uses only the IDSP disease case count data without any exogenous input to estimate the current case count. Contrary to this, the Tweet-ARMAX model, TweetX-LSTMConcat, TweetX-LSTMAverage, and Multivariate-LSTM models use twitter and IDSP data to estimate the current case counts. The twitter data in these models are used as exogenous input to incorporate the information embedded within the trends of tweets. Figure 5 presents line graphs of one-step-ahead case count estimation of dengue cases using CaseCount-ARMA, Tweet-ARMAX, CaseCount-LSTM, TweetX-LSTMConcat, TweetX-LSTMAverage, and Multivariate-LSTM models during the validation period.

We used the Normalized Root Mean Square Error (NRMSE) [51] to compare the case count estimations by each model during the evaluation period. The NRMSE is calculated using equation 11, where RMSE is the root mean square error given by equation 12, and y_{max} and y_{min} respectively are the maximum and minimum number of cases during the validation period. In equation 12 y_t and $\hat{y}_t$ represent observed and predicted dengue cases, respectively.

$$NRMSE = \frac{RMSE}{y_{max} - y_{min}} \tag{11}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(\hat{y}_t - y_t)^2} \tag{12}$$

TABLE I: Performance comparison of CaseCount-ARMA and Tweet-ARMAX for 1-step ahead estimation of dengue case counts

Models	Normalized RMSE
CaseCount-ARMA	0.294
Tweet-ARMAX	0.220
CaseCount-LSTM	0.215
TweetX-LSTMConcat	0.179
TweetX-LSTMAverage	0.217
Multivariate-LSTM	0.184

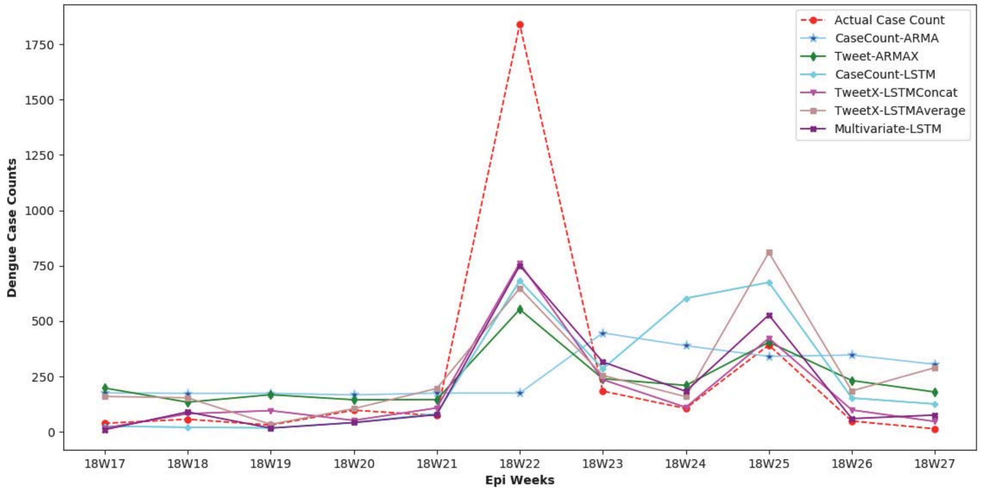

Fig. 5: The graph showing relationship between the actual dengue cases and the case counts estimated by various models during the validation period

Table I presents the performance of all the models for case count estimation in terms of NRMSE. The NRMSE values from table I clearly show that the LSTM-based models outperform the auto regressive moving average models as the NRMSE values are smaller for LSTM-based models in comparison to the auto regressive moving average models. If we compare the two auto regressive moving average models (`CaseCount-ARMA` and `Tweet-ARMAX`), we observe that the NRMSE value is significantly smaller in the case of the `Tweet-ARMAX` model in comparison to the `CaseCount-ARMA` model, indicating that the tweet volumes per week as an exogenous variable can improve the performance for estimating the disease case count. Further, we observe from table I that `TweetX-LSTMConcat` has the least NRMSE value suggesting that it is the best performing model for predicting dengue cases while the `CaseCount-ARMA` model having the highest NRMSE value indicates worst performing model for predicting dengue cases. Moreover, the LSTM models having inputs both the historic dengue case count data and the dengue tweet volumes perform better than the models having input only the historic dengue case count data. Thus, we can infer that the `Twitter` data can be instrumental in determining the dengue cases and can act as a complementary source of information for early reporting of dengue outbreaks. The early estimation of disease case counts can help Govt. health agencies to take suitable precautionary and preventive measures to prevent a disease to become epidemic or pandemic.

VI. Conclusion and Future Work

Estimation of disease incidences is a major task in identifying and monitoring of the disease outbreak, tracking infection rate, and reporting them to public health authorities responsible for intervention. In this paper, we have presented an approach for the monitoring and estimation of disease cases using neural network-based LSTM models with input data from disease-related tweets volumes and the disease case count data reported through the Govt. health agencies. Capturing the trends of the disease-related social media data is useful for forecasting. However, insufficient and inconsistent coverage of the disease in social media resources during the outbreak may adversely affect the nowcasting or forecasting accuracy. We observed that social media data could be used as a complementary source of information for nowcasting or estimating the current disease case counts. There is still scope of improvement in nowcasting in the scenarios when little or inadequate coverage of the disease by social media. We can supplement the model with additional information such as climatic conditions like temperature, humidity, rainfall, and other seasonal features to enhance the disease incidence accuracy.

References

[1] M. Abulaish, M. A. Parwez, and Jahiruddin, "Disease: A biomedical text analytics system for disease symptom extraction and characterization," *Journal of Biomedical Informatics*, vol. 100, no. 103324, pp. 1–23, 2019. [Online]. Available: https://doi.org/10.1016/j.jbi.2019.103324

[2] M. M. Wagner, A. W. Moore, and R. M. Aryel, *Handbook of biosurveillance*. Elsevier, 2011.

[3] J. S. Lombardo and D. L. Buckeridge, *Disease surveillance: a public health informatics approach*. John Wiley & Sons, 2012.

[4] S. C. Chaintoutis, C. I. Dovas, M. Papanastassopoulou, S. Gewehr, K. Danis, C. Beck, S. Lecollinet, V. Antalis, S. Kalaitzopoulou, T. Panagiotopoulos *et al.*, "Evaluation of a west nile virus surveillance and early warning system in greece, based on domestic pigeons," *Comparative Immunology, Microbiology and Infectious Diseases*, vol. 37, no. 2, pp. 131–141, 2014.

[5] J. S. Lombardo, H. Burkom, and J. Pavlin, "Essence ii and the framework for evaluating syndromic surveillance systems," *Morbidity and Mortality Weekly Report*, pp. 159–165, 2004.

[6] G. Eysenbach, "Infodemiology: tracking flu-related searches on the web for syndromic surveillance," in *AMIA Annual Symposium Proceedings*, vol. 2006. American Medical Informatics Association, 2006, p. 244.

[7] C. Chew and G. Eysenbach, "Pandemics in the age of twitter: content analysis of tweets during the 2009 h1n1 outbreak," *PloS one*, vol. 5, no. 11, p. e14118, 2010.

[8] A. Culotta, "Towards detecting influenza epidemics by analyzing twitter messages," in *Proceedings of the first workshop on social media analytics*. Acm, 2010, pp. 115–122.

[9] H. Achrekar, A. Gandhe, R. Lazarus, S.-H. Yu, and B. Liu, "Twitter improves seasonal influenza prediction." in *Healthinf*, 2012, pp. 61–70.

[10] S. Ghosh, P. Chakraborty, E. Cohn, J. S. Brownstein, and N. Ramakrishnan, "Characterizing diseases from unstructured text: A vocabulary driven word2vec approach," in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM, 2016, pp. 1129–1138.

[11] S. S. Amrith, "Health in india since independence," in *History, historians and development policy*. Manchester University Press, 2020.

[12] P. Nsubuga, M. E. White, S. B. Thacker, M. A. Anderson, S. B. Blount, C. V. Broome, T. M. Chiller, V. Espitia, R. Imtiaz, D. Sosin *et al.*, "Public health surveillance: a tool for targeting and monitoring interventions," *Disease control priorities in developing countries*, vol. 2, pp. 997–1018, 2006.

[13] F. Rohart, G. J. Milinovich, S. M. Avril, K.-A. Lê Cao, S. Tong, and W. Hu, "Disease surveillance based on internet-based linear models: an australian case study of previously unmodeled infection diseases," *Scientific Reports*, vol. 6, 2016.

[14] N. Thapen, D. Simmie, C. Hankin, and J. Gillard, "Defender: Detecting and forecasting epidemics using novel data-analytics for enhanced response," *PloS one*, vol. 11, no. 5, p. e0155417, 2016.

[15] B. M. Althouse, S. V. Scarpino, L. A. Meyers, J. W. Ayers, M. Bargsten, J. Baumbach, J. S. Brownstein, L. Castro, H. Clapham, D. A. Cummings *et al.*, "Enhancing disease surveillance with novel data streams: challenges and opportunities," *EPJ data science*, vol. 4, no. 1, p. 17, 2015.

[16] D. Paolotti, A. Carnahan, V. Colizza, K. Eames, J. Edmunds, G. Gomes, C. Koppeschaar, M. Rehn, R. Smallenburg, C. Turbelin *et al.*, "Web-based participatory surveillance of infectious diseases: the influenzanet participatory surveillance experience," *Clinical Microbiology and Infection*, vol. 20, no. 1, pp. 17–21, 2014.

[17] N. Generous, G. Fairchild, A. Deshpande, S. Y. Del Valle, and R. Priedhorsky, "Global disease monitoring and forecasting with wikipedia," *PloS Computational Biology*, vol. 10, no. 11, p. e1003892, 2014.

[18] P. M. Polgreen, Y. Chen, D. M. Pennock, F. D. Nelson, and R. A. Weinstein, "Using internet searches for influenza surveillance," *Clinical infectious diseases*, vol. 47, no. 11, pp. 1443–1448, 2008.

[19] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data," *Nature*, vol. 457, no. 7232, p. 1012, 2009.

[20] J. Gomide, A. Veloso, W. Meira Jr, V. Almeida, F. Benevenuto, F. Ferraz, and M. Teixeira, "Dengue surveillance based on a computational model of spatio-temporal locality of twitter," in *Proceedings of the 3rd international web science conference*. ACM, 2011, pp. 1–8.

[21] R. T. Gluskin, M. A. Johansson, M. Santillana, and J. S. Brownstein, "Evaluation of internet-based dengue query data: Google dengue trends," *PLoS neglected tropical diseases*, vol. 8, no. 2, p. e2713, 2014.

[22] Z. Li, T. Liu, G. Zhu, H. Lin, Y. Zhang, J. He, A. Deng, Z. Peng, J. Xiao, S. Rutherford *et al.*, "Dengue baidu search index data can improve the prediction of local dengue epidemic: A case study in guangzhou, china," *PLOS neglected tropical diseases*, vol. 11, no. 3, p. e0005354, 2017.

[23] R. Chunara, J. R. Andrews, and J. S. Brownstein, "Social and news media enable estimation of epidemiological patterns early in the 2010 haitian cholera outbreak," *The American journal of tropical medicine and hygiene*, vol. 86, no. 1, pp. 39–45, 2012.

[24] A. J. Ocampo, R. Chunara, and J. S. Brownstein, "Using search queries for malaria surveillance, thailand," *Malaria journal*, vol. 12, no. 1, p. 390, 2013.

[25] M. Odlum and S. Yoon, "What can we learn about the ebola outbreak from tweets?" *American journal of infection control*, vol. 43, no. 6, pp. 563–571, 2015.

[26] A. J. Lazard, E. Scheinfeld, J. M. Bernhardt, G. B. Wilcox, and M. Suran, "Detecting themes of public concern: a text mining analysis of the centers for disease control and prevention's ebola live twitter chat," *American journal of infection control*, vol. 43, no. 10, pp. 1109–1111, 2015.

[27] M. Miller, T. Banerjee, R. Muppalla, W. Romine, and A. Sheth, "What are people tweeting about zika? an exploratory study concerning its symptoms, treatment, transmission, and prevention," *JMIR public health and surveillance*, vol. 3, no. 2, p. e38, 2017.

[28] S. F. McGough, J. S. Brownstein, J. B. Hawkins, and M. Santillana, "Forecasting zika incidence in the 2016 latin america outbreak combining traditional disease surveillance with search, social media, and news report data," *PLoS neglected tropical diseases*, vol. 11, no. 1, p. e0005295, 2017.

[29] J. Eschler, Z. Dehlawi, and W. Pratt, "Self-characterized illness phase and information needs of participants in an online cancer forum," in *Ninth International AAAI Conference on Web and Social Media*, 2015, pp. 101–109.

[30] M. J. Paul, R. W. White, and E. Horvitz, "Search and breast cancer: On episodic shifts of attention over life histories of an illness," *ACM Transactions on the Web (TWEB)*, vol. 10, no. 2, p. 13, 2016.

[31] H. Dai, B. R. Lee, and J. Hao, "Predicting asthma prevalence by linking social media data and traditional surveys," *The ANNALS of the American Academy of Political and Social Science*, vol. 669, no. 1, pp. 75–92, 2017.

[32] Y. Liu, Q. Mei, D. A. Hanauer, K. Zheng, and J. M. Lee, "Use of social media in the diabetes community: an exploratory analysis of diabetes-related tweets," *JMIR diabetes*, vol. 1, no. 2, p. e4, 2016.

[33] J. Han, X. Tian, G. Yu, and F. He, "Disclosure pattern of self-labeled people living with hiv/aids on chinese social networking site: An exploratory study," *Cyberpsychology, Behavior, and Social Networking*, vol. 19, no. 8, pp. 516–523, 2016.

[34] S. D. Young, W. Yu, and W. Wang, "Toward automating hiv identification: machine learning for rapid identification of hiv-related social media data," *Journal of acquired immune deficiency syndromes (1999)*, vol. 74, no. Suppl 2, p. S128, 2017.

[35] Z. Tufekci, "Big questions for social media big data: Representativeness, validity and other methodological pitfalls," in *Eighth International AAAI Conference on Weblogs and Social Media*, 2014, pp. 505–514.

[36] T. Harford, "Big data: A big mistake?" *Significance*, vol. 11, no. 5, pp. 14–19, 2014.

[37] D. Lazer, R. Kennedy, G. King, and A. Vespignani, "The parable of google flu: traps in big data analysis," *Science*, vol. 343, no. 6176, pp. 1203–1205, 2014.

[38] J. Mowery, "Twitter influenza surveillance: Quantifying seasonal misdiagnosis patterns and their impact on surveillance estimates," *Online journal of public health informatics*, vol. 8, no. 3, pp. 1–19, 2016.

[39] R. Chunara, L. E. Wisk, and E. R. Weitzman, "Denominator issues for personally generated data in population health monitoring," *American journal of preventive medicine*, vol. 52, no. 4, pp. 549–553, 2017.

[40] R. Allard, "Use of time-series analysis in infectious disease surveillance." *Bulletin of the World Health Organization*, vol. 76, no. 4, pp. 327–333, 1998.

[41] P. M. Luz, B. V. Mendes, C. T. Codeço, C. J. Struchiner, and A. P. Galvani, "Time series analysis of dengue incidence in rio de janeiro, brazil," *The American journal of tropical medicine and hygiene*, vol. 79, no. 6, pp. 933–939, 2008.

[42] M. A. Parwez, M. Abulaish, and M. Fazil, "Drcove: An augmented word representation approach using distributional and relational context," in *Proceedings of the 16th International Conference on Natural Language Processing (ICON), Hyderabad, India*, December 18-21, 2019, pp. 1–10.

[43] S. Siami-Namini, N. Tavakoli, and A. S. Namin, "A comparison of arima and lstm in forecasting time series," in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2018, pp. 1394–1401.

[44] ——, "A comparative analysis of forecasting financial time series using arima, lstm, and bilstm," *arXiv preprint arXiv:1911.09512*, 2019.

[45] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: forecasting and control (5th Edt.)*. John Wiley & Sons, 2015.

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

[47] Z. C. Lipton, J. Berkowitz, and C. Elkan, "A critical review of recurrent neural networks for sequence learning," *arXiv preprint arXiv:1506.00019*, 2015.

[48] M. A. Parwez, M. Abulaish, and Jahiruddin, "Multi-label classification of microblogging texts using convolution neural network," *IEEE Access*, vol. 7, pp. 68 678–68 691, 2019.

[49] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *ICML*, 2010.

[50] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[51] M. V. Shcherbakov, A. Brebels, N. L. Shcherbakova, A. P. Tyukov, T. A. Janovsky, and V. A. Kamaev, "A survey of forecast error measures," *World Applied Sciences Journal*, vol. 24, no. 24, pp. 171–176, 2013.