Health and Technology
https://doi.org/10.1007/s12553-019-00309-4

ORIGINAL PAPER

Predicting the spread of influenza epidemics by analyzing twitter messages

Soheila Molaei [1] · Mohammad Khansari [1] · Hadi Veisi [1] · Mostafa Salehi [1]

Received: 22 December 2018 / Accepted: 25 February 2019

Abstract

Seasonal influenza epidemics affect millions of people with respiratory illnesses and cause 250,000 to 500,000 deaths worldwide each year. Rapidly predicting the outbreak of epidemics leads to an earlier detection and control. In this study, we predicted an influenza-like illness (ILI) based on social media data derived from Twitter. Tweets and patients do not always have a linear correlation; therefore, we employed nonlinear methods including autoregressive with exogenous inputs (ARX), autoregressive-moving-average with exogenous inputs (ARMAX), nonlinear autoregressive exogenous (NARX), deep multilayer perceptron (DeepMLP), and a convolutional neural network (CNN). Two new features employed to significantly reduce the prediction errors are products of the tweets and Centers for Disease Control and Prevention (CDC) data and of the tweets and Google data. Furthermore, we introduced a new method based on entropy that decreased the errors as well as time complexity. Among the available methods and features, the best results were obtained with the newly developed features in the deep neural network methods and the entropy-based method that decreased the mean average error by up to 25%. The entropy method also reduced the time complexity. Applying the above-mentioned methods to the Twitter datasets from 2009 to 2010 and 2011–2014 revealed that the ILI outbreak can be predicted 2–4 weeks earlier than by the CDC.

Keywords Contagious disease prediction · Influenza · Twitter · Deep neural network · Nonlinear method · Entropy

1 Introduction

Seasonal influenza epidemics lead to practically 500,000 deaths every year worldwide [1, 2]. Between 1918 and 1920, Spanish influenza led to 50–100 million deaths globally [3]. Therefore, reducing the spread of diseases such as H1N1-type influenza is of foremost importance. One approach is to track and predict the disease dispersion process in a population. Research has shown that several such diseases can be controlled if diagnosed in advance [4, 5].

The Center for Disease Control and Prevention (CDC) [6] that controls influenza-like illnesses (ILI) by collecting data, prepares its reports on a weekly, monthly, and annual basis. The CDC is among the most credible organizations; however, approximately 1–2 weeks are required following a disease diagnosis by physicians for the publication of its reports. Ideally, governments must be informed instantly so that a possible outbreak of diseases such as influenza can be prevented.

Various observation systems have been recommended to control the relevant health behaviors and recognize the activity of the influenza virus. Some of these systems include telephone triages for monitoring ILI [7] and the correlation between over-the-counter pharmaceutical sales and volume of patients with ILI [8]. In this respect, Google [9] recommends a tool to evaluate an influenza outbreak process based on the recorded growth rate of influenza-related searches.

In addition to Google, online social networks that have become increasingly prominent in recent years, have also been used to predict respiratory syndrome [10], headaches [11], and

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s12553-019-00309-4) contains supplementary material, which is available to authorized users.

✉ Mohammad Khansari
m.khansari@ut.ac.ir

Soheila Molaei
soheila.molaei@ut.ac.ir

Hadi Veisi
h.veisi@ut.ac.ir

Mostafa Salehi
mostafa_salehi@ut.ac.ir

[1] Faculty of New Sciences and Technologies, University of Tehran, End of North Kargar Street, Tehran, Iran

Published online: 21 March 2019

epidemic processes [12]. Twitter, for example, is a social network that allows users to send and read short 140-character text messages called 'tweets' [13]. The basic concept is to monitor health trends such as the outbreak of contagious diseases like influenza with the aid of such online networks. For this purpose, first, we require a real dataset of Twitter. The data can be collected from the website of Twitter by writing a web application (and using the Twitter application programming interfaces (API)) [14]. (Fig. 1) shows the general process of predicting the influenza epidemics by collecting several types of data and analyzing them. The CDC [6], which controls ILI by collecting data, releases its report on a weekly, monthly, and annual basis. This center is among the most reliable organizations, but there is a gap of 1–2 weeks between the disease diagnosis by physicians and publication of the reports. However, governments must be provided the necessary information instantly to be able to prevent an outbreak possibility of diseases such as influenza.

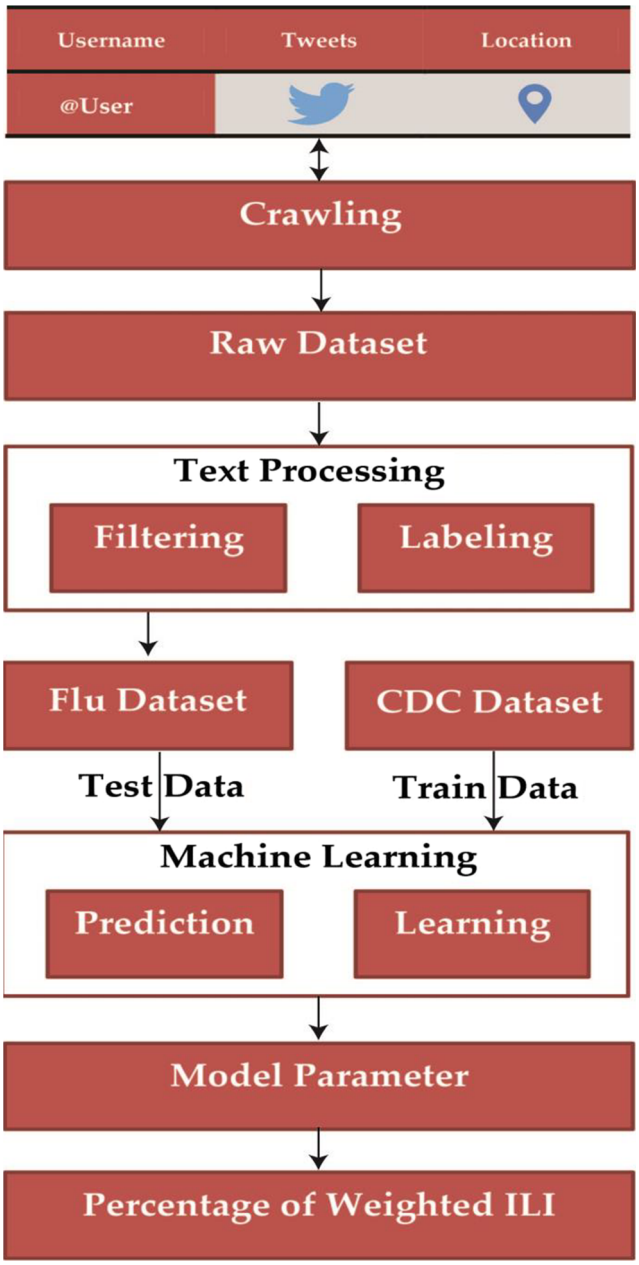

Fig. 1 Process for predicting influenza epidemics

Fig. 1 shows the general process of data collection and prediction based on previous studies. First, Twitter is crawled by using its search API at regular intervals, and tweets with flu-related keywords ("Flu," "H1N1," and "Swine Flu") are collected. Subsequently, using text processing techniques, misleading tweets that suppress the actual data (e.g., tweets with words such as "vaccination" and "flu shot") are removed using a filtering technique. Moreover, the tweets are tagged as "patient" or "healthy" based on whether they are indicative of a flu event, by using a labeling technique. Tweets are nonnumeric; therefore, it is required to extract the data from the tweets by tagging each of them as "patient" or "healthy." Using the labeling techniques, a series of tweets is considered as a training set and used for learning. This learning assigns weights to the words and thereby, determines whether the "patient" or "healthy" tag is appropriate. The number of tweets that are tagged as "patient" in n days (in this study, $n = 7$) is considered as the input data. Employing this data and the reference data (from the CDC), the necessary information is obtained, parameters are estimated, and the percentage of patients is predicted. Finally, to validate the proposed method, the predictions are compared with the weighted ILI (i.e., the percentage of patients reported by the CDC) and the associated error is calculated.

Generally, for predicting disease outbreaks, methods such as machine learning, data mining, text mining [15, 16], and probabilistic graphical models such as support vector machines (SVMs), regression [17–20], and conditional random field (CRF) [21] have been used. Tagging every tweet as "patient" or "healthy" requires a considerable amount of tagged data that can, in turn, tag other tweets by machine learning [17–21]. The existing methods [22] predict diseases on a weekly basis using tweets of t weeks, ahead of the CDC reports by approximately 1–2 weeks.

The main contribution of this study is the use of methods such as autoregressive with exogenous inputs (ARX), autoregressive-moving-average with exogenous inputs (ARMAX), nonlinear autoregressive exogenous (NARX), and deep neural networks to enhance the learning and prediction modules shown in (Fig. 1). In contrast to the existing solutions that primarily use linear prediction, we applied nonlinear methods. The proposed methods predict the outbreak of diseases 2–4 weeks prior to the CDC.

The main contributions of this work are as follows:

- Introduction of two new features that are products of tweets and the CDC data and of tweets and the Google data. These were used in the proposed methods and led to

an earlier prediction of the percentage of patients with a lower percentage error compared to previous methods. They were also able to simply indicate the nonlinear interaction between tweets and the CDC.
- Application and comparison of various nonlinear methods such as NARX and deep neural network models for prediction.
- Introduction of a new method based on entropy for obtaining the CDC data that also decreased the errors and time complexity.

The remainder of this paper is organized as follows: In Related Works, we discuss the other research projects that are related to our research. In Methods, we present a brief introduction to data collection and our proposed methods. In Experiment and Results, we explain the results of the implementation of these proposed methods. Finally, in Conclusions, we present the conclusions of our study and provide future directions for this research.

2 Related works

Recently, several related studies have been performed employing social networks such as Facebook, Twitter, Flickr, LinkedIn, Wikipedia, and YouTube. [23, 24]. Asur et al. [25] demonstrated how social media content such as tweets can be utilized to forecast future outcomes of box-office revenues for movies in advance of their release. Motoyama et al. [26] used Twitter data to detect outages in several widely used web services such as Amazon, Gmail, Google, PayPal, Netflix, YouTube, Facebook, Wikipedia, and Flickr. Similarly, Mislove et al. [27] presented the moods of the population of the United States from Twitter messages describing the variation in their mood over a 24-h period as well as the days of a week.

Efforts have been made for utilizing the Twitter data for gauging public interest or determining the concern for health-related issues such as toothache [28], cardiac arrest and resuscitation [29], allergies, obesity and sleeplessness [30], and dengue [31] that is an infectious disease transmitted by a type of mosquito found in Brazil and is the major cause of death in tropical and subtropical areas such as Brazil.

In contrast, Signorini et al. [32] examined the use of the information embedded in the Twitter stream to track the rapidly evolving public sentiment regarding H1N1 or swine flu. Their results showed that they could establish a distinct relationship between the Twitter data and the epidemic curve of the 2009 H1N1 outbreak, both nationally and geographically. Similarly, Chew and Eysenbach [33] evaluated Twitter for monitoring the public perception of the 2009 H1N1 pandemic. In addition, Lampos and Cristianini [34] conducted a research on H1N1 influenza in England by analyzing the tweets recorded within 24 weeks. In this research, statements that warned about disease symptoms were analyzed and finally converted into an influenza score.

In comparison, Culotta [19] studied the misleading messages on Twitter and showed that only a small percentage of influenza-related keywords can predict future influenza rates. Aramaki et al. [35] imposed a support vector machine method on tweets to predict the prevalence rate of influenza in Japan. Their experiments demonstrated that their proposed method detects influenza epidemics up to a 0.89 correlation.

In another instance, Achrekar et al. [15, 22] provided an instant tool called "SNEFT" [36] to control the influenza trend via the social network. This tool mainly created a tool to control influenza through information on Twitter and Facebook. To this end, a list of keywords such as influenza and H1N1, as well as disease-related messages and user information (such as location and time), was generated based on the data collected from the tweets, and then applied to a framework for disease prediction.

In addition to Twitter, Google has been used to predict the outbreak of diseases. Following the study by Ginsberg et al. [37], Google search is used to elicit the linear correlation between the number of searches and number of patients who have consulted physicians.

Google Flu Trends [9] tracks the influenza-related searches on Google and thereby predicts the outbreak of diseases 1–2 weeks earlier than the CDC. FluTrackers [38] has been following up the flu disease since 2006. Similarly, Flu Survey [39] is an online system to track influenza in the United Kingdom. Unlike the CDC, these sites collect data directly from the public, rather than from hospitals. Each week, registered users input their experience of flu on the website.

Yahoo! has also been used for a similar purpose. Polgreen et al. [40] proved that the frequency of flu-related queries on the Yahoo! search engine is related to influenza and mortality in the United States. Similarly, Hulth et al. [41] determined the ILI rate in Sweden based on the queries made by users on a Swedish medical website.

The problem with the existing methods is that they only predict the current week trends. Any attempt at predicting the occurrence of diseases the next week increases the percentage of error. To solve this problem, the proposed methods use data that are 1–2 weeks old. Therefore, these methods can predict 1–2 weeks ahead with a lower percentage of error. The lower percentage of error has always been very important. To achieve this, we have introduced two new features that support the prediction. We can see that there exists a remarkable nonlinear interaction between the tweets and CDC data. We also proposed a new method based on entropy, which reduced the errors and obviated the problem of the neural network method (time complexity).

3 Methods

3.1 Data analysis

Fig. 1 shows that for predicting influenza, we first require data. In this study, we used two datasets.

3.1.1 Dataset 1-Twitter 2009–10

Achrekar et al. [15, 22] searched the tweets posted by users between October 15, 2009 and October 31, 2010 containing influenza-related keywords, and collected America-related tweets. During this period, there were as many as 4.7 million tweets from 1.5 million unique Twitter users. They stated that because of a power outage on their data collection site, no data were collected between January 18, 2010 and January 20, 2010. Subsequently, retweets and the next tweets from the same user were deleted. From the remaining tweets, the ones with words such as "vaccination" and "flu shot" were deleted.

3.1.2 Dataset 2-Twitter 2011–2014

We used the data collected by Paul et al. [42] from November 27, 2011 to April 5, 2014, and this data is related to the United States. They used the Twitter influenza surveillance system developed by Lamb et al. [17, 43].

Table 1 Variables used in paper

Variables	Description
$Y(t)$	Percentage of patients due to CDC in week t
$\hat{Y}$	Predicted CDC Data
$\overline{Y}$	Mean of CDC Data
U_t	Influenza-related tweets
U_g	Google Flu Trends
$e(t)$	A sequence of independent random variables
a, b, c, d	Numerical coefficients of the regression
$u(t)$	Input
P	Linear Subspace Matrix
Q	Nonlinear Subspace Matrix
r	Mean of $u(t)$
L	Linear Correlation Vector
a_k	Scaling coefficient (scalar)
b_k	Scaling Dilation (scalar)
c_k	Wavelet Dilation (scalar)
d_k	Wavelet coefficient (scalar)
e_k	Wavelet translation (vector)
h_k	Scaling translation (scalar)
$C(t)$	The Coefficient of obtaining Y in week (t)

3.1.3 Google flu trends

This influenza surveillance system estimates the current infection rates based on the volume of Google searches for a select number of influenza-related queries [9]. We gathered the Google data in the same time interval as that of the Twitter data.

3.2 Influenza Prediction

The proposed methods are described in this section. The variables used in this paper are listed in (Table 1).

3.3 Regression models

The general regression model is expressed in (1). We can generate several different models from it by modifying each of the parameters.

$$Y(t) = \sum_{i}^{m} a_i Y(t-i) + \sum_{j}^{n} b_j U_t(t-j) + \sum_{l}^{o} d_l U_g(t-l) + \sum_{k}^{p} e(t) \qquad (1)$$

3.3.1 Autoregressive with exogenous inputs (ARX)

Autoregressive (AR) is used generally for time-varying processes. The autoregressive model shows the dependence of the output on previous values. When an AR model uses exogenous inputs, it is converted into an ARX model [44]. A linear regression model with an exogenous input is expressed in (2). We call this equation *ARX (In1, In2, Out),* where *In1* and *In2* denote $U_t(t-j)$ and $U_g(t-l)$, respectively and *Out* defines $Y(t-i)$. This model uses tweets and Google in week t and the CDC data from the previous weeks to predict the percentage of patients in week t.

$$Y(t) = \sum_{i=2}^{m} a_i Y(t-i) + \sum_{j=0}^{n} b_j U_t(t-j) + \sum_{l=0}^{o} d_l U_g(t-l) + e(t) \qquad (2)$$

Additionally, we can add products of the tweets and CDC data or of the tweets and Google data as input, as shown in (3).

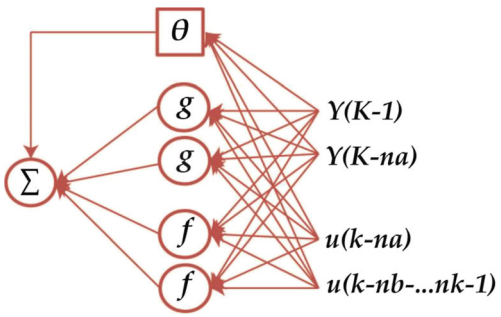

Fig. 2 Nonlinear autoregressive with the wavelet function

We call this equation *ARX_Product (In1, In2, In3, Out),* where *In1, In2,* and *In3* denote $U_t(t-j)$, $U_g(t-l)$, and $U_1(t-l)$, respectively and *Out* defines $Y(t-i)$.

$$Y(t) = \sum_{i=2}^{m} a_i Y(t-i) + \sum_{j=0}^{n} b_j U_t(t-j) + \sum_{l=0}^{o} d_l U_g(t-l) + \sum_{r=2}^{q} c_r U_1(t-r) + e(t) \quad (3)$$

$$where \rightarrow U_1 = Y^* U_t \text{ or } U_1 = U_g{}^* U_t$$

This is the regression equation with tweets and the CDC (or Google) data and the products of the tweets and CDC (or Google) data as its exogenous inputs.

3.3.2 Autoregressive-moving-average with exogenous inputs (ARMAX)

Equation (4) represents an ARMAX model. We name this equation as *ARMAX (In1, In2, Out),* where *In1* and *In2* denote $U_t(t-j)$ and $U_g(t-l)$, respectively and *Out* defines $Y(t-k)$.

$$Y(t) = \sum_{i=1}^{m} a_i Y(t-i) + \sum_{j=0}^{n} b_j U_t(t-j) + \sum_{l=0}^{o} d_l U_g(t-l) + \sum_{k=1}^{p} e(k) \quad (4)$$

Equation (5) defines an ARMAX with exogenous inputs in the form of the CDC (or Google) data and products of the tweets and CDC (or Google) data. We call this equation *ARMAX_Product (In1, In2, In3, Out),* where *In1, In2 and In3* refer to $U_t(t-j)$, $U_g(t-l)$, and $U_1(t-r)$, respectively and *Out* represents $Y(t-i)$.

$$Y(t) = \sum_{i=2}^{m} a_i Y(t-i) + \sum_{j=2}^{n} b_j U_t(t-j) + \sum_{l=0}^{o} d_l U_g(t-l) + \sum_{r=2}^{q} c_r U_1(t-r) + \sum_{k=1}^{p} e(k) \quad (5)$$

3.3.3 Nonlinear autoregressive exogenous (NARX)

This is a nonlinear ARX model that features dynamic systems and is a combination of linear and nonlinear functions. The output of this function is defined below [45, 46].

$$Y(t) = \mathrm{F}(\mathrm{u}(\mathrm{t})), Y(t) = \mathrm{F}\left(\sum_{i=2}^{m} Y(t-i), \sum_{j=0}^{n} U_t(t-j), \sum_{l=0}^{o} U_g(t-l)\right) + \sum_{l=1}^{p} e(l) \quad (6)$$

Equation (6) defines a nonlinear autoregressive with exogenous inputs where F is a nonlinear function. Here, function F is intended to be a tree-partition and wavelet function. We call this equation as *NARX (In1, In2, Out),* where *In1* and *In2* refer to $U_t(t-j)$, $U_g(t-l)$, respectively and *Out* defines $Y(t-i)$.

WaveNet is defined as $x = F(u)$ and is as follows:

$$F(u) = (u-r)PL + a_1 f(b_1((u-r)Q-h_k)) + \ldots + a_k f(b_k((u-r)Q-h_k)) + d_1 g(c_1((u-r)Q-e_1)) + d_k g(c_k((u-r)Q-e_k)) + d$$

$$f(u) = \exp\left(-0.5uu'\right), g(u) = \left(\dim(u)-u'u\right)\exp\left(-0.5u'u\right) \quad (7)$$

In (7), f is a scaling function and g is a wavelet function [45, 46]. Fig. 2 depicts a nonlinear autoregressive with the wavelet function.

The tree-partition function is as follows:

$$F(u) = d + uL + (1,u)C_k \quad (8)$$

We built a tree with 2^{j-1} vertices. Here, each node is assigned to group P so that for any P, we have a function F called the piecewise function, and C_k is computed by the least-square error algorithm. We called this equation as *NARX_Product (In1, In2, Out),* where *In1, In2,* and *In3* refer to $U_t(t-j)$, $U_g(t-l)$, and $U_1(t-j)$, respectively, and *Out* represents $Y(t-i)$.

$$Y(t) = \mathrm{F}\left(\sum_{i=2}^{m} Y(t-i), \sum_{j=0}^{n} U_t(t-j), \sum_{l=0}^{o} U_g(t-l), \sum_{j=2}^{q} U_1(t-j)\right) + e(t) \quad (9)$$

Equation (9) presents a nonlinear autoregressive with the exogenous inputs of tweets and the CDC (or Google) data and products of the tweets and CDC (or Google) data.

3.4 Deep multi-layer perceptron (deep MLP)

One of the primary methods in a deep neural network is MLP (Fig. 3). A feedforward neural network, or MLP [47], is a computational model that processes information through a series of interconnected computational nodes. These computational nodes are grouped into layers and are associated with one another using weighted connections. The nodes of the layers are called units (or neurons) and transform the data by means of non-linear operations to create a decision boundary for the input by projecting it into space where it becomes

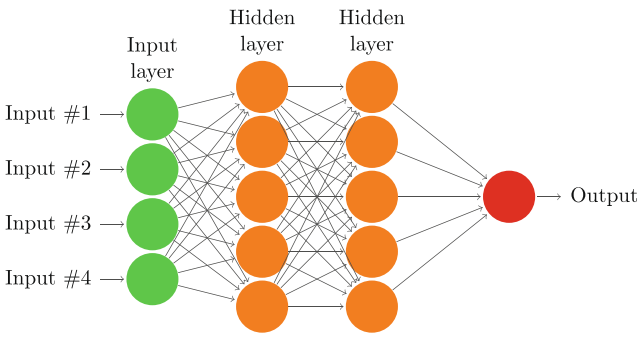

Fig. 3 Deep multi-layer perceptron

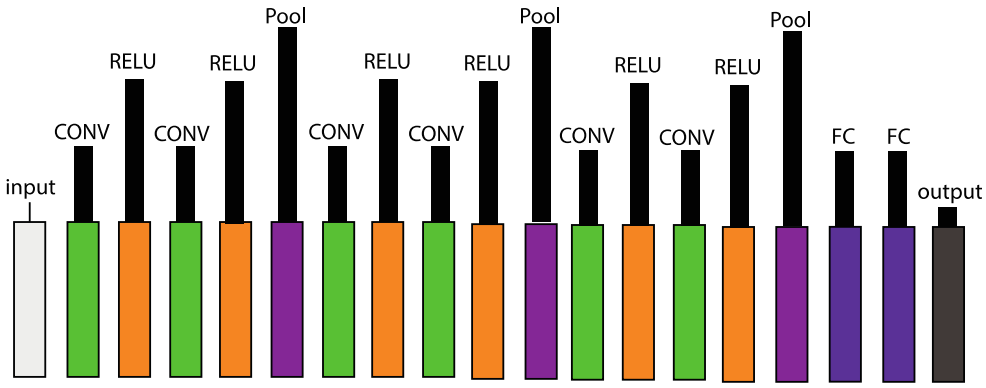

Fig. 4 Convolutional Neural Networks

linearly separable [50]. We used two MLP layers with relu activation function and regularized using dropout. Sigmoid used as an activation function at the output layer that demonstrates the CDC data prediction in the current week.

In this method, we have used a multi-layer perceptron neural network with three layers. We have also used a ReLu function in the hidden layer and a sigmoid function in the output layer. We named this model as *DeepMLP (In1, In2, In3, Out),* where *In1, In2,* and *In3* refer to U_t, U_g, and U_1, respectively and *Out* represents *Y.*

3.5 Convolutional neural networks (CNN)

Convolutional layers consist of a set of learnable filters (Fig. 4). During the forward pass, each filter convolves across the width and height of the input volume to produce a 2-dimensional activation map. These activation maps will stack along the depth dimension and produce the output volume. More generally, a convolutional layer requires four hyper-parameters including, the number of filters *K*, their spatial extent *F*, the stride with which they are applied *S*, and the amount of zero padding on the borders of the input, *P.* A convolutional layer accepts a tensor of *size W1* × *H1* × *D1* and produces an output volume of size *W2* × *H2* × *D2*. The number of parameters in each filter is *F* × *F* × *D1*, for a total of *(F* × *F* × *D1)* × *K* weights and *K* biases. In the output tensor, each d–th slice of the output of size *W2* × *H2* is the result of performing a valid convolution of the d–th filter over the input tensor with a stride of *S* and then offsetting the result by d–th bias.

CNNs are designed in such a way, that they can take into account the spatial structure of the input. They were inspired by mice visual system.

3.6 Entropy-based method

In this section, we introduce a new method based on the Shannon entropy [48].

$$E(t) = -\sum_{i=1}^{u(t)} p_i \log(p_i) = -\sum_{i=1}^{u(t)} \frac{1}{i} \log\left(\frac{1}{i}\right) \tag{10}$$

where the sum is over all the states (the definition by Shannon included a constant that has been removed here). Let E represent the entropy of a set of probabilities $p_i, \ldots, p_{u(t)}$. $p_{u(t)}$ is the probability of choosing one tweet from all the infected tweets in week t, and therefore we calculated the probability for each tweet based on (10). In this formula, we can replace $u(t)$ with U_t *and* U_g as inputs. For using multiple inputs, we can sum the normalized inputs (same range) and replace with $u(t)$. For example, if we want to use tweets and the Google data together, we should normalize them in the same range and replace their sum with $u(t)$. We obtained the coefficient as expressed in (11).

$$C(t) = \frac{E(t) \times E(t-1) \times \ldots \times E(t-k+1)}{E(t-1) \times E(t-2) \times \ldots \times E(t-k)} = \frac{\prod_{i=0}^{k-1} E(t-i)}{\prod_{i=0}^{k-1} E(t-i-1)}$$
$$Y(t) = C(t) \times Y(t-k) \tag{11}$$

Therefore, we can predict k steps ahead, implying that we can obtain the CDC prediction in week t from week *t-k*.

3.6.1 Solved example

Assuming that we have the data as given in Table 2, then we can predict the CDC estimates of weeks 2 and 3 as follows:

$$E(1) = -\sum_{i=1}^{45057} \frac{1}{i} \log\left(\frac{1}{i}\right) = -24.9026, E(2)$$
$$= -\sum_{i=1}^{43102} \frac{1}{i} \log\left(\frac{1}{i}\right) = -24.6965, E(3)$$
$$= -\sum_{i=1}^{27512} \frac{1}{i} \log\left(\frac{1}{i}\right) = -22.6597$$

$$k = 1 : C(2) = \frac{\prod_{i=0}^{0} E(t-i)}{\prod_{i=0}^{0} E(t-i-1)} = \frac{E(2)}{E(1)} = 0.9917, Y(2)$$
$$= 0.9917^{*}7.688 = 7.62$$

Table 2 Solved example

Week	Patient Tweets	CDC
1	45,057	7.688
2	43,102	?
3	27,512	?

$$k = 2 : C(3) = \frac{\prod_{i=0}^{1} E(t-i)}{\prod_{i=0}^{1} E(t-i-1)} = \frac{E(3) \times E(2)}{E(2) \times E(1)}$$
$$= 0.9175 \times 0.9917 = 0.9099, Y(3) = 0.9099 \times 7.688$$
$$= 6.99$$

4 Experiment and results

In this study, we used two datasets. To achieve more accurate answers, we adopted the five-fold cross-validation (CV) approach. With aim of proposed methods, the disease outbreak can be predicted 2–4 weeks earlier than by the CDC. Determining how many weeks ahead the outbreak of diseases can be predicted depends on the input parameter. It can be set so that the model can predict two to four weeks earlier than by the CDC.

4.1 Data

Dataset 1 is from 2009 to 2010. This dataset is for one season, referred as Season 1. Dataset 2 is from 2011 to 2014, and contains three seasons (Seasons 2–4). Season 2 is from 2011 to 2012, Season 3 is from 2012 to 2013, and Season 4 is from 2013 to 2014.

4.2 Evaluated metrics and notations

Some comparison measures and notations are described in this section.

4.2.1 Metrics

For comparative purposes, we calculated the mean absolute error (MAE) and root-mean-square error (RMSE). Calculations of MAE and RMSE percentages are defined in (12) and (13), respectively.

$$\mathrm{MAE}\% = 100.\frac{1}{\mathrm{N}} \sum_{k=1}^{N} |\hat{Y}(\mathrm{k}) - Y(\mathrm{k})| \tag{12}$$

$$\mathrm{RMSE}\% = \frac{100.\sqrt{\frac{1}{\mathrm{N}} \sum_{\mathrm{k}=1}^{\mathrm{N}} \left[\hat{Y}(\mathrm{k}) - Y(\mathrm{k})\right]^2}}{Y_{max} - Y_{min}} \tag{13}$$

4.2.2 Notations

Some notations used in this paper are as follows:

Arx (In1, In2, Out): Here we should replace In1 with the tweets, In2 with the Google data, and finally Out with the CDC data. For example, Arx (tweet,0,CDC) refers to (2), where In1 = tweets and In2 = 0, and Out = CDC data.

Arx_Product (In1, In2, In3, Out): In1, In2, and In3 denote the tweets, Google data, products of the tweets and CDC data or of tweets and the Google data.

Arx_Product (tweets, 0, tweets*CDC, CDC) corresponds to In1 = tweets, In2 = 0, In3 = product of the tweets and CDC data, and Out = CDC data.

This type of notation is also applied for ARMAX (In1, In2, Out), ARMAX_Product (In1, In2, In3, Out), NARX (In1, In2, Out), NARX_Product (In1, In2, In3, Out), and NN(In1, In2, In3, Out).

Entropy (In1, In2, In3): In1, In2, and In3 denote the tweets, Google data, and products of the tweets and Google data. However, when we used multiple inputs, we summed them as one input and replaced with $u(t)$ in (10).

4.3 Experimental results

We applied various inputs in the proposed methods including ARX, ARMAX, NARX, NN, and Entropy. ARX with tweets

Table 3 Results of the ARX method

Methods	Dataset 1		Dataset 2						Average Error of Dataset 2	
	RMSE	MAE	RMSE			MAE				
	Season 1 (09–10)		Season 2 (11–12)	Season 3 (12–13)	Season 4 (13–14)	Season 2 (11–12)	Season 3 (12–13)	Season 4 (13–14)	RMSE	MAE
ARX(Tweet,0,0)	8.12	32.02	23.20	9.73	17.47	22.95	28.13	36.17	16.8	29.08
ARX(0,0,CDC)	9.14	39.32	13.38	10.90	12.41	14.68	30.22	26.85	12.23	23.91
ARX(0,Google,0)	35.75	182.3	20.72	10.54	14.84	26.27	31.84	49.97	15.36	36.02
ARX(Tweet,0,CDC)	6.97	31.24	14.22	9.56	10.42	14.50	25.70	22.35	11.4	20.85
ARX(Tweet,Google,0)	14.86	54.86	18.69	10.54	10.54	19.04	31.84	27.69	13.25	26.19

Table 4 Results of the ARX_Product method

Methods	Dataset 1		Dataset 2						Average Error of Dataset 2	
	RMSE	MAE	RMSE			MAE				
	Season 1 (09–10)		Season 2 (11–12)	Season 3 (12–13)	Season 4 (13–14)	Season 2 (11–12)	Season 3 (12–13)	Season 4 (13–14)	RMSE	MAE
*ARX_Product (Tweet,0,Tweet*CDC,CDC)*	6.56	24.69	10.00	8.92	10.20	12.61	24.66	22.49	9.70	19.92
*ARX_Product (Tweet,Google,Google*Tweet,0)*	6.77	28.94	14.13	8.42	8.99	16.20	24.71	19.01	10.51	19.97

as input (ARX [Tweet, 0, 0]) was considered as a base method for comparison with other methods because other similar studies [22, 42] used ARX as their method. We used the same data for the same periods that these studies [22, 42]) used.

4.3.1 ARX

(Table 3) shows that tweets as input provide better predictions than the CDC data in Dataset 1. CDC predicts more precisely in Dataset 2, therefore, we decided to use the CDC data and tweets simultaneously in the proposed methods. In both the datasets, using the tweets and CDC data together decreases the errors.

In Dataset 1, Google alone increases the errors, whereas in Dataset 2, the errors of three seasons reduce. However, in Dataset 1, using the tweets and Google data together as input did not reduce the errors compared with only tweets as input. In contrast, in Dataset 2, using the tweets and CDC data and the tweets and Google data together decreases the MAE by 8.2% and 2.8%, respectively.

4.3.2 ARX_product

According to (Table 4), using the product of the tweets and CDC data as input decreases the MAE by 7.3% in Dataset 1 and by 9.1% in Dataset 2. Furthermore, another new feature is the product of the tweets and Google data that decreases the MAE by 3% in Dataset 1 and by 9.1% in Dataset 2.

4.3.3 ARMAX

(Table 5) shows that in the ARMAX model, using the tweets and CDC data together reduces the errors. However, the most accurate results are with the Google data in Dataset 1 and the Google data and tweets together in Dataset 2. The Google data improves the MAE by 4.2% in Dataset 1 and by 14.3% in Dataset 2. The MAE is decreased by 13.5% in Dataset 2 by using the Google data and tweets together. Based on (4), this model integrates the noises (e) in each step apparently making it more accurate than ARX.

4.3.4 ARMAX_product

According to (Table 6), by using the products of the tweets and CDC data and of the tweets and Google data in the ARMAX model, the MAE reduces by 3.1% and 4%, respectively in Dataset 1 and by 15.9% in Dataset 2.

Table 5 Results of the ARMAX method

Methods	Dataset 1		Dataset 2						Average Error of Dataset 2	
	RMSE	MAE	RMSE			MAE				
	Season 1 (09–10)		Season 2 (11–12)	Season 3 (12–13)	Season 4 (13–14)	Season 2 (11–12)	Season 3 (12–13)	Season 4 (13–14)	RMSE	MAE
ARMAX(Tweet,0,0)	6.61	34.51	12.48	6.44	9.47	13.75	18.34	22.46	9.46	18.18
ARMAX (0,0,CDC)	5.99	30.63	19.58	7.11	7.05	25.26	19.52	15.99	11.24	20.25
ARMAX (0,Google,0)	5.77	27.80	9.17	5.70	5.59	11.10	16.07	17.10	6.82	14.75
ARMAX (Tweet,0,CDC)	6.32	29.23	16.00	5.83	6.71	20.74	18.19	13.36	9.51	17.43
ARMAX (Tweet,Google,0)	14.86	54.86	9.43	7.68	5.00	10.90	21.47	14.28	6.34	15.55

Table 6 Results of the ARMAX_Product method

Methods	Dataset 1		Dataset 2						Average Error of Dataset 2	
	RMSE	MAE	RMSE			MAE				
	Season 1 (09–10)		Season 2 (11–12)	Season 3 (12–13)	Season 4 (13–14)	Season 2 (11–12)	Season 3 (12–13)	Season 4 (13–14)	RMSE	MAE
*ARMAX_Product (Tweet,0,Tweet*CDC,CDC)*	6.12	28.85	8.53	5.11	4.60	9.94	16.43	13.05	6.08	13.14
*ARMAX_Product (Tweet,Google,Google*Tweet,0)*	6.07	27.99	7.63	6.75	5.10	8.51	17.03	13.76	6.49	13.10

4.3.5 NARX

In the NARX model, tweets, CDC data, Google data, products of tweets and the CDC data, and products of the Google data and tweets were tested as inputs, and all of them reduce the errors (see (Table 7)). Considering (Table 7), the best results are obtained with using the Google data and tweets together in Dataset 1. In Dataset 2, the best answers are achieved with the tweets and CDC data as well as with the tweets and Google data that reduce the MAE by 9.1% and 6.7%, respectively.

The results in (Table 7) reveal that the nonlinear methods are effective.

4.3.6 NARX_product

According to (Table 8), the NARX model with the newly introduced features remarkably decreases the errors. Using products of the tweets and CDC data decreases the RMSE and MAE by 2.1% and 3.6%, respectively in Dataset 1 and by 11% and 15.3%, respectively in Dataset 2. Furthermore, the other new feature of products of tweets and Google data decreases the RMSE and MAE by 2.9% and 7.1%, respectively in Dataset 1 and by 12.5% and 22.9%, respectively in Dataset 2. The products of the tweets and CDC data as well as of the tweets and Google data are more nonlinear in nature and, therefore, the predicted errors are significantly smaller.

4.3.7 Deep MLP

(Table 9) illustrates that the best responses are for the neural network model with all the inputs compared with the remaining described models. The best answers are obtained with the products of the tweets and CDC data and of the tweets and Google data.

The products of the tweets and CDC data decrease the RMSE and MAE by 3.8% and 10.4%, respectively in Dataset 1 and by 12.7% and 21.8%, respectively in Dataset 2. By using the products of the tweets and Google data as input, the RMSE and MAE are reduced by 4.3% and 13%, respectively in Dataset 1 and by 13.6% and 24.6%, respectively in Dataset 2.

4.3.8 CNN

As shown in (Table 10), in CNN model, all inputs have acceptable results but the products of the tweets and CDC decreases the RMSE and MAE error by 3.12% and 4.43% respectively which is the best result compared to other inputs results.

Table 7 Results of the NARX method

Methods	Dataset 1		Dataset 2						Average Error of Dataset 2	
	RMSE	MAE	RMSE			MAE				
	Season 1 (09–10)		Season 2 (11–12)	Season 3 (12–13)	Season 4 (13–14)	Season 2 (11–12)	Season 3 (12–13)	Season 4 (13–14)	RMSE	MAE
NARX(Tweet,0,0)	8.06	38.31	23.58	9.46	11.64	26.37	26.72	29.32	14.89	27.47
NARX(0,0,CDC)	7.75	38.42	17.74	8.25	11.29	17.45	25.92	27.93	12.42	23.76
NARX(0,Google,0)	6.50	34.30	19.13	9.56	15.46	20.09	37.81	35.51	14.71	31.13
NARX(Tweet,0,CDC)	7.04	34.33	6.47	8.25	11.29	5.88	25.92	27.93	8.67	19.91
NARX(Tweet,Google,0)	5.62	29.57	15.46	8.77	10.47	15.84	27.43	23.79	11.56	22.35

Table 8 Results of the NARX_Product method

Methods	Dataset 1		Dataset 2						Average Error of Dataset 2	
	RMSE	MAE	RMSE			MAE				
	Season 1 (09–10)		Season 2 (11–12)	Season 3 (12–13)	Season 4 (13–14)	Season 2 (11–12)	Season 3 (12–13)	Season 4 (13–14)	RMSE	MAE
*NARX_Product (Tweet,0,Tweet*CDC,CDC)*	6.01	28.40	5.42	8.65	3.16	5.64	27.68	7.87	5.74	13.27
*NARX_Product (Tweet,Google,Google*Tweet,0)*	5.14	24.84	8.94	1.34	2.40	7.75	4.29	6.41	4.22	6.15

4.3.9 Entropy-based method

The results of this method are similar to those of the neural network method, but obtaining these results requires less time. Using the tweets, Google data, and products of the tweets and Google data have approximately the same effect as shown in (Table 11). This method can reduce the time complexity in addition to lowering the error because it does not require the time associated with the learning process.

4.3.10 Discussion

Among the linear models, the best result was related to the *Arx* model, with the inputs of tweets, CDC and multiplications of tweets and CDC as well as tweets, Google and multiplications of tweets and Google with the average RMSE error of 6.08% and 6.49% respectively. But in nonlinear methods, when multiplications of the inputs were included, the error was significantly reduced. In the *NARX* method, this error reaches 5.74% and 4.22% respectively, and in the *DeepMLP* method, it is 4.8% and 3.14% respectively, which is a very small error. While in the entropy method, that is nonlinear in nature, the error was at its lowest, which is 3.21%.

As described in the analysis section (Appendix), using Kendall's tau correlation [49] method, there is a nonlinearity between input data (Twitter) and output (CDC), which is why non-linear methods are more appropriate to solve this problem. To illustrate the correctness of our approach, linear methods such as *ARX* and *ARMAX* have been used for comparison with non-linear methods like *NARX*, *DeepMLP, CNN* and *Entropy.*

Also, choosing appropriate input parameters is a significant matter. The input parameter could be the number of tweets, but we thought about using other inputs such as the CDC, which is the source of information gathering, and Google data, which predicts flu. These added inputs help our models to converge faster to actual results.

There is a gap of few weeks between the disease diagnosis by physicians and publication of the reports by CDC so for prediction we could use existed CDC data in *(t-k)* which we can predict *k* step ahead. Therefore, based on results, we tried to show that these data are also useful in predicting. Eventually, for emphasizing on nonlinearity of data, we introduced two non-linear features, namely products of the tweets and CDC data and products of the tweets and Google data.

Figures 5 and 6 show that by using these features, the conclusive results significantly improve, and in all the methods they decrease the errors compared with other inputs. In these figures, *ARX* being the method having been used in other similar studies [22, 42], has been applied as the base method for comparison.

Table 9 Results of Deep MLP method

Methods	Dataset 1		Dataset 2						Average Error of Dataset 2	
	RMSE	MAE	RMSE			MAE				
	Season 1 (09–10)		Season 2 (11–12)	Season 3 (12–13)	Season 4 (13–14)	Season 2 (11–12)	Season 3 (12–13)	Season 4 (13–14)	RMSE	MAE
DeepMLP(Tweet, 0, 0, 0)	6.06	28.13	14.40	6.49	8.9	14.38	18.01	22.45	9.93	18.28
DeepMLP (Tweet, 0, 0, CDC)	5.78	27.67	10.14	4.89	7.21	7.11	9.85	9.05	7.41	8.67
DeepMLP(Tweet, Google, 0, 0)	5.62	28.57	6.08	5.02	7.29	6.00	9.99	9.45	6.13	8.48
*DeepMLP (Tweet, 0, Tweet*CDC, CDC)*	4.24	21.55	4.45	4.65	3.16	4.67	9.09	7.87	4.08	7.21
*DeepMLP (Tweet, Google, Tweet*Google, 0)*	3.79	18.93	5.74	1.35	2.35	5.00	3.99	4.31	3.14	4.43

Table 10 Results of the CNN method

Methods	Dataset 1		Dataset 2						Average Error of Dataset 2	
	RMSE	MAE	RMSE			MAE				
	Season 1 (09–10)		Season 2 (11–12)	Season 3 (12–13)	Season 4 (13–14)	Season 2 (11–12)	Season 3 (12–13)	Season 4 (13–14)	RMSE	MAE
CNN(Tweet, 0, 0, 0)	5.96	27.09	4.5	4.8	5	7.03	7.31	8.42	4.76	7.58
CNN (Tweet, 0, 0, CDC)	6.00	28.00	5.36	4.19	5.73	7.17	7.32	8.26	5.09	7.58
CNN (Tweet, Google, 0, 0)	5.65	28.50	4.6	4.6	5.8	7.11	8.01	8.73	5	7.95
*CNN (Tweet, 0, Tweet*CDC, CDC)*	4.54	22.01	3.43	3.00	2.95	4.44	4.67	4.20	3.12	4.43
*CNN (Tweet, Google, Tweet*Google, 0)*	3.69	18.06	4.12	3.92	4.13	6.4	5.3	6.43	4.05	6.04

As we go from linear methods to non–linear methods, the results will get better. Recalling, among the non–linear methods, the deep neural network methods yielded the best results, particularly with the two new introduced features. However, we intend to present a method with less time complexity. The entropy-based method satisfied this objective and also improved the results effectively.

The reason that the added features had a positive impact on reducing the percentage of error is mentioned below:

- When products of these two features are used, it implies that the shared characteristics of these two features are used. Multiplying the tweets and CDC or Google data shows the significant nonlinear interaction between these two inputs and makes them more effective. Figures 7 and 8 displays that the prediction with deep learning methods is in excellent agreement to the CDC data prediction.

The reasons that the Deep MLP method yields a better result among the other methods could be the following:

- Neural Networks are a rich combination of the switching states of the neurons to produce a nonlinear input–output mapping function. Furthermore, as shown in (Appendix), the tweets and CDC data exhibit a nonlinear relationship, so that using neural networks appears to be more logical. Moreover, NNs with several sets of weighted connections between the neurons can learn to represent various ranges of the input–output mapping problems [47]. Recalling, the best results were obtained by multiplying the tweets and CDC data and the tweets and Google data in the NNs. As mentioned earlier, a DeepMLP is nonlinear in nature, and multiplying the tweets and CDC or Google data represents the notable nonlinear interaction between them. Thus, using these inputs in a powerful nonlinear method (NN) can be expected to yield the best results.

The reasons why CNNs have satisfying answers are as follows:

- Extracting as much information as possible from the available data sets is crucial to creating an effective solution and for small datasets extracting feature could be helpful.
- One of the most impressive characteristics of CNN is feature learning. CNN can learn relevant features.

The reasons that the entropy method provides acceptable results besides decreasing the calculation time could be the following:

Table 11 Results of the entropy-based method

Methods	Dataset 1		Dataset 2						Average Error of Dataset 2	
	RMSE	MAE	RMSE			MAE				
	Season 1 (09–10)		Season 2 (11–12)	Season 3 (12–13)	Season 4 (13–14)	Season 2 (11–12)	Season 3 (12–13)	Season 4 (13–14)	RMSE	MAE
Entropy(Tweet, 0, 0)	3.92	19.08	4.02	2.93	3.01	5.09	4.93	5.66	3.32	5.22
Entropy (Tweet, Google, 0)	4.27	19.57	5.37	3.28	4.08	5.97	4.98	7.25	4.23	6.06
*Entropy (Tweet, Google, Tweet*Google)*	3.86	18.01	4.55	2.12	2.96	5.12	4.67	4.80	3.21	4.86

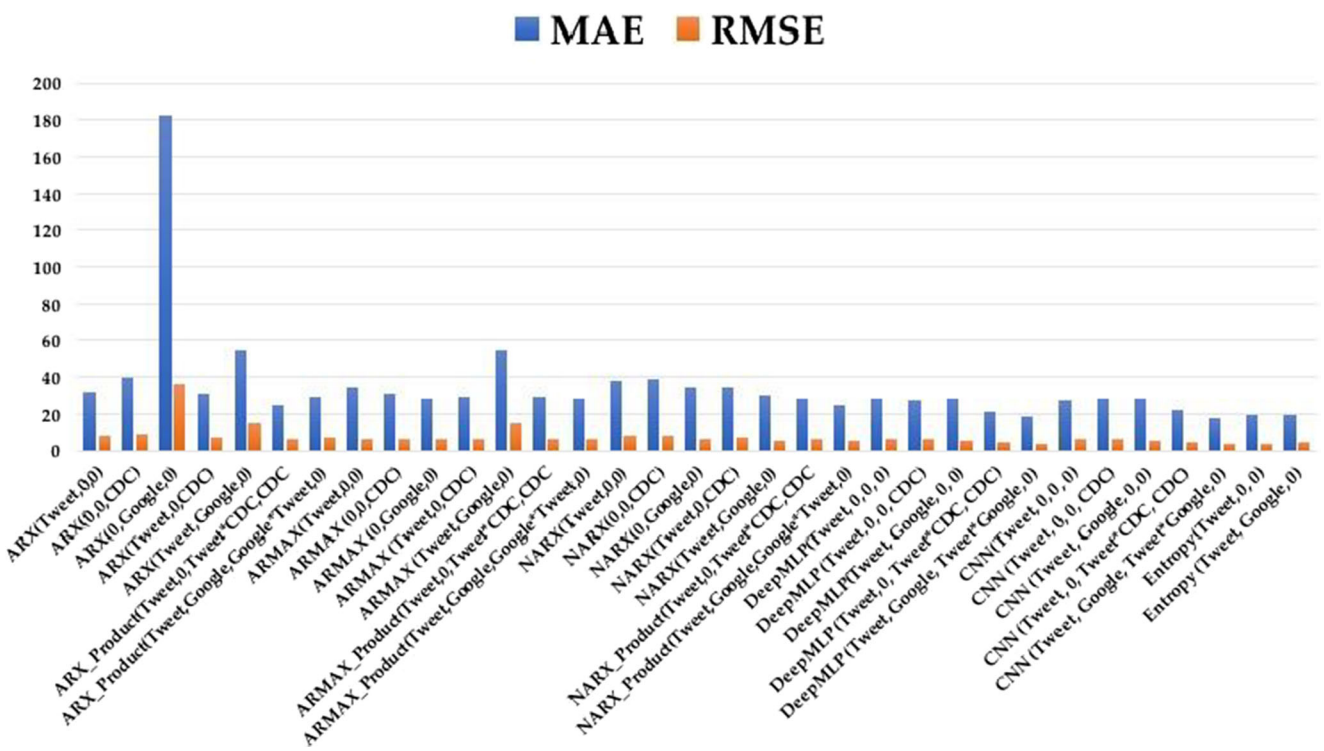

Fig. 5 MAE errors of the proposed methods with Dataset 1

- With the entropy-based method, we showed that increasing (or decreasing) slope of the tweets and CDC data is not linear, i.e., the increasing trend of the tweets is faster than of the CDC data, and we should control this. Therefore, a linear prediction is not always appropriate, and the entropy causes the obtained coefficient to move toward nonlinearity and controls the speed of the increasing trend of the tweets. For more clarity, if we have "43,102" patient

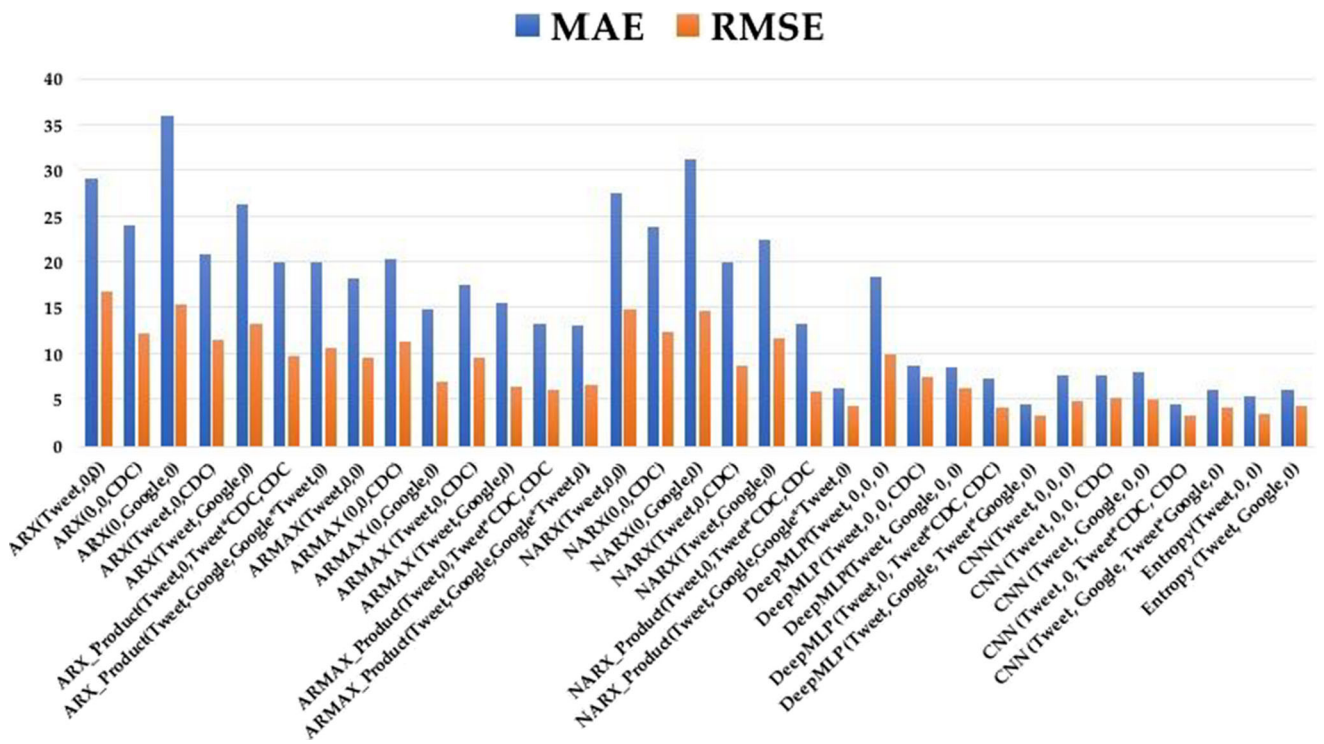

Fig. 6 Errors of the proposed methods with Dataset 2

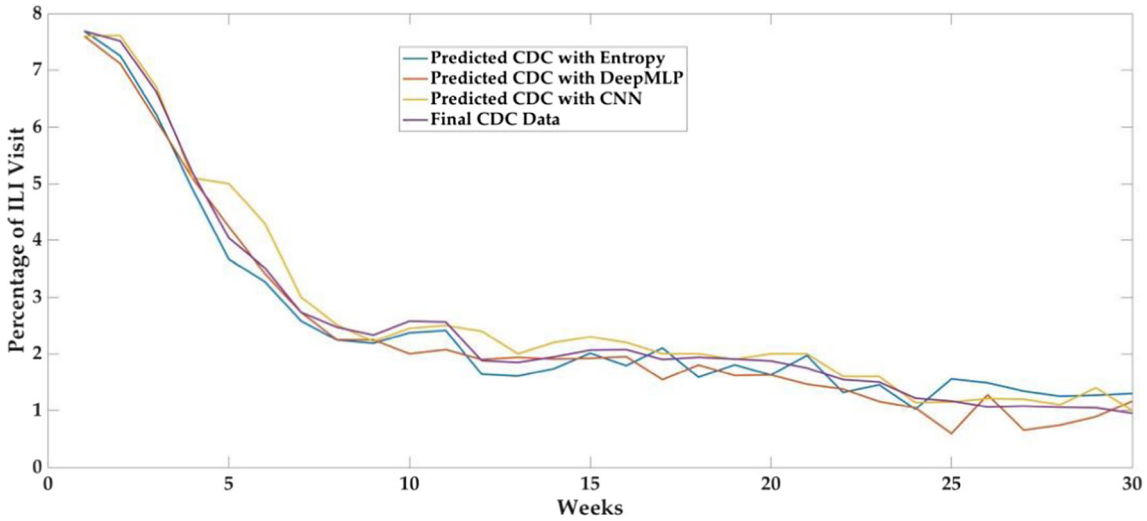

Fig. 7 Results of Proposed Methods with Dataset 1

tweets in week (1) and "45,057" in week (2), in a linear assumption the increasing slope is 45,057/43102 = 1.045, but with the entropy method, this trend is $\sum_{i=1}^{45057} \frac{1}{i}\log\left(\frac{1}{i}\right) / \sum_{i=1}^{43102} \frac{1}{i}\log\left(\frac{1}{i}\right) = 1.008$. Fig. 9 illustrates the increasing trends of the tweets with and without entropy and the real trend of the CDC data.

1. The Twitter data provides a real-time assessment of the flu epidemic (i.e. the availability of Twitter data in week t in the prediction of physician visits also in week t while we assume that the CDC data has 2 weeks of lag in data reporting and aggregation. Hence, the multiplying of twitter and CDC data can capture the flu trend.
2. There were no events at the time period in which the experiment is conducted that might perturb the Twitter data. Hence, Twitter data reflects the true scope of the epidemic.
3. In all methods, by changing k in *t-k*, we can predict *k* steps ahead which is related to the size of the dataset. Also for the bigger dataset, we can use more hidden layers in Deep Networks.

For more clarity, the Summary of results are shown in Table 12.

4.4 Complexity

Multi-layer perceptron neural network: this method uses the backpropagation technique for training. Therefore, suppose

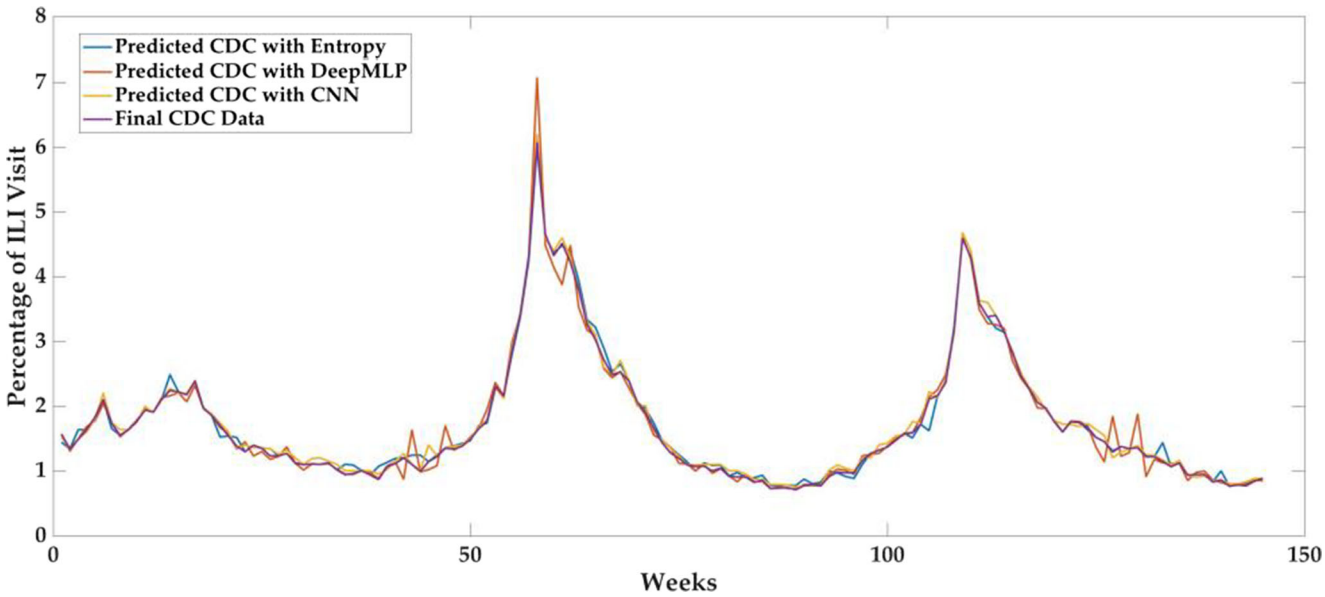

Fig. 8 Results of Proposed Methods with Dataset 2

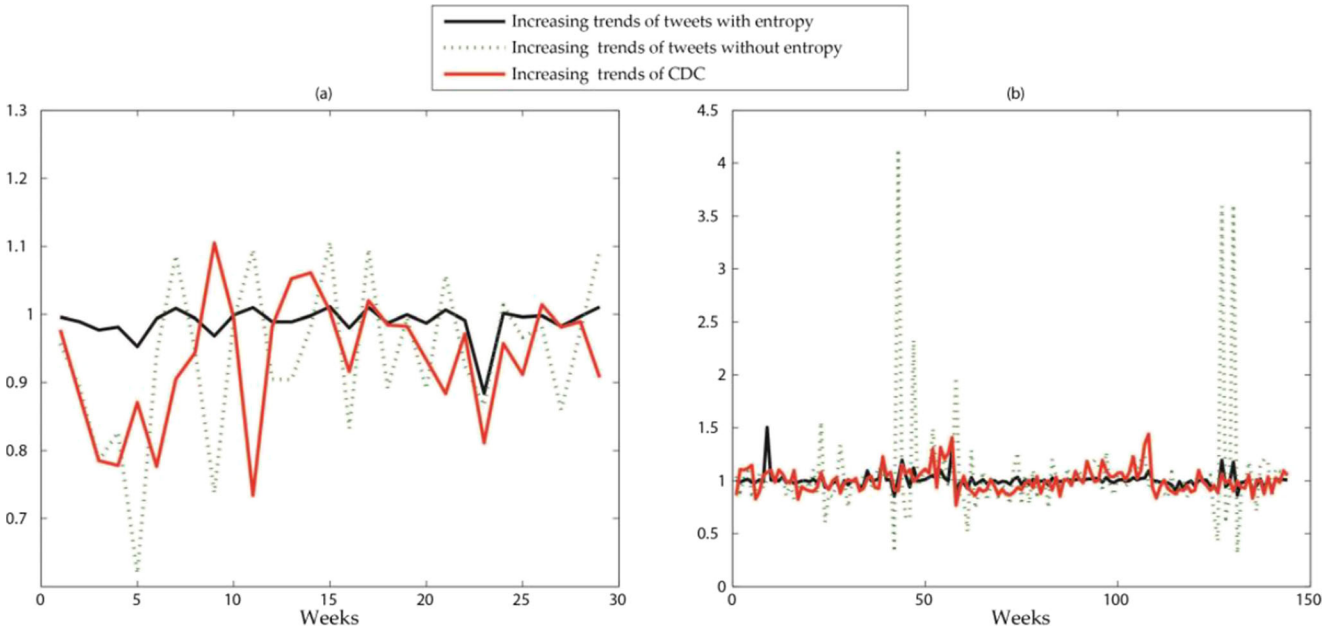

Fig. 9 **a** is an increasing trend of Dataset 1 and **b** is an increasing trend of Dataset 2. Entropy controlled the sudden changes in the increasing trend of the tweets

there are n training samples, m features, k hidden layers, each containing h neurons and o output neurons. The time complexity of the backpropagation is O ($n. m. h^k. o. i$), where i is the number of iterations. Backpropagation has a high time complexity.

Entropy: for the prediction with this method, each input requires the value of prior inputs so that the time complexity is O (n).

5 Conclusions

In this study, the prediction of an ILI was conducted with the data derived from Twitter. Owing to the variable nature of the data and the fact that these data are not always linear, the topic of nonlinear methods emerged. Therefore, we made efforts to shift from linear methods to nonlinear methods and compared the results. To predict the percentages of patients visited by physicians, we used machine learning methods such as autoregressive with exogenous inputs (ARX), autoregressive-moving-average with exogenous inputs (ARMAX), nonlinear autoregressive exogenous (NARX), and deep neural network (DeepMLP and CNN) time series model.

We also introduced two new features, namely, products of the tweets and CDC data and of the tweets and Google data that had a significant impact on reducing the errors in the linear and nonlinear models. Furthermore, we suggested an entropy-based method that improved the results and decreased the time complexity significantly.

Table 12 Summary of Proposed Methods Results

Methods	Average Error of Dataset 1		Average Error of Dataset 2	
	RMSE	MAE	RMSE	MAE
ARX(Tweet,0,CDC)	6.97	31.24	11.4	20.85
*ARX_Product (Tweet,0,Tweet*CDC,CDC)*	6.56	24.69	9.70	19.92
ARMAX (0,Google,0)	5.77	27.80	6.82	14.75
*ARMAX_Product (Tweet,Google,Google*Tweet,0)*	6.07	27.99	6.49	13.10
NARX(Tweet,0,CDC)	7.04	34.33	8.67	19.91
NARX(Tweet,Google,0)	5.62	29.57	11.56	22.35
*NARX_Product (Tweet,Google,Google*Tweet,0)*	5.14	24.84	4.22	6.15
*DeepMLP (Tweet, Google, Tweet*Google, 0)*	3.79	18.93	3.14	4.43
*CNN (Tweet, 0, Tweet*CDC, CDC)*	4.54	22.01	3.12	4.43
*Entropy (Tweet, Google, Tweet*Google)*	3.86	18.01	3.21	4.86

Among the proposed methods, the best results were obtained from the deep neural network methods and the entropy-based method that decreased the mean average error by up to 25%. Applying these methods to the Twitter datasets from 2009 to 2010 and 2011–2014 showed that the disease outbreak can be predicted 2–4 weeks earlier than by the CDC.

In the future, it could be possible to recognize infected persons via social networks. This could be achieved based on the number of people connected to an infected person, and it could be possible to prevent infection and the occurrence of diseases.

Acknowledgements We would like to thank MJ. Paul, M. Dredze, D. Broniatowski for allowing us to use their collecting Twitter dataset.

Compliance with ethical standards

Conflicts of interest Soheila Molaei, Mohammad Khansari, Hadi Veisi and Mostafa Salehi declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Duda K. Flu Deaths Per Year. about Heal. 2016. Available from: https://www.verywell.com/flu-deaths-per-year-770503. Accessed 2017.
2. Guo P, Zhang J, Wang L, Yang S, Luo G, Deng C, et al. Monitoring seasonal influenza epidemics by using internet search data with an ensemble penalized regression model. Sci Rep. 2017;7:46469. Available from: http://www.nature.com/articles/srep46469. Accessed 3 March 2019.
3. Morens DM, Fauci AS. The 1918 influenza pandemic: insights for the 21st century. J Infect Dis. 2007;195:1018–28.
4. Paul MJ, Dredze M. A model for mining public health topics from Twitter. Health (Irvine Calif). 2012;11:16. Available from: http://www.cs.jhu.edu/~mpaul/files/2011.tech.twitter_health.pdf. Accessed 13 March 2019.
5. Chen L, Hossain KSMT, Butler P, Ramakrishnan N, Prakash BA. Flu Gone Viral: Syndromic Surveillance of Flu on Twitter Using Temporal Topic Models. Proc - IEEE Int Conf Data Mining, ICDM. 2015. p. 755–60.
6. Centers for Disease Control and Prevention. a Wkly. Influ. Surveill. Rep. 2009. Available from: http://www.cdc.gov/. Accessed 3 March 2019.
7. Yih WK, Teates KS, Abrams A, Kleinman K, Kulldorff M, Pinner R, et al. Telephone triage service data for detection of influenza-like illness. PLoS One. 2009;4.
8. Liu TY, Sanders JL, Tsui FC, Espino JU, Dato VM, Suyama J. Association of Over-The-Counter Pharmaceutical Sales with Influenza-Like-Illnesses to Patient Volume in an Urgent Care Setting. PLoS One. 2013;8.
9. Google Flu Trends. Available from: http://www.google.org/flutrends/us/data.txt. Accessed 2017.
10. Shin S-Y, Seo D-W, An J, Kwak H, Kim S-H, Gwack J, et al. High correlation of Middle East respiratory syndrome spread with Google search and Twitter trends in Korea. Sci Rep. 2016;6:32920. Available from: http://www.nature.com/articles/srep32920. Accessed 3 March 2019.
11. Milojević S. Revisiting the connection between Solar eruptions and primary headaches and migraines using Twitter. Sci Rep. 2016;6:39769. Available from: http://www.nature.com/articles/srep39769. Accessed 3 March 2019.
12. Tizzoni M, Sun K, Benusiglio D, Karsai M, Perra N. The Scaling of Human Contacts and Epidemic Processes in Metapopulation Networks. Sci Rep. 2015;5:15111. Available from: http://www.nature.com/articles/srep15111. Accessed 3 March 2019.
13. Posting a Tweet. 2017. Available from: https://support.twitter.com/articles/15367. Accessed 2017.
14. Twitter Developer Platform(API). 2014. Available from: https://developer.twitter.com. Accessed 3 March 2019.
15. Achrekar H, Lazarus R, Park WC. Predicting Flu Trends using Twitter Data. IEEE Infocom. 2011;702–7.
16. Lee K. Real-time disease surveillance using twitter data: demonstration on flu and cancer. KDD'13. 2013;1474–7.
17. Lamb A, Paul MJ, Dredze M. Separating fact from fear: tracking flu infections on Twitter. Conf North Am Chapter Assoc Comput Linguist Hum Lang Technol. 2013;
18. Sadilek A, Kautz H, Silenzio V. Modeling spread of disease from social interactions. Int AAAI Conf Weblogs Soc Media. 2012.
19. Culotta A. Towards detecting influenza epidemics by analyzing Twitter messages. Proc First Work Soc Media Anal - SOMA '10. New York, New York, USA: ACM Press; 2010;115–22. Available from: http://portal.acm.org/citation.cfm?doid=1964858.1964874. Accessed 3 March 2019.
20. Bodnar T, Salathé M. Validating Models for Disease Detection Using Twitter Regression on Tweet Count. Proc 22nd Int Conf World Wide Web companion. 2013;699–702.
21. Peng H-K, Zhu J, Piao D, Yan R, Zhang Y. Retweet modeling using conditional random fields. 2011 IEEE 11th Int Conf Data Min Work. IEEE; 2011;336–43. Available from: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6137399. Accessed 3 March 2019.
22. Achrekar H, Gandhe A, Lazarus R, Yu S, Liu B. Online Social Networks Flu Trend Tracker: A Novel Sensory Approach to Predict Flu Trends. Springer. 2013. p. 353–68. Available from: http://link.springer.com/10.1007/978-3-642-38256-7_24. Accessed 3 March 2019.
23. Caverlee J, Webb S, Tech G. A Large-Scale Study of MySpace : Observations and Implications for Online Social Networks. Proc from 2nd Int Conf Weblogs Soc Media AAAI. 2008;
24. Gauvin W, Ribeiro B, Towsley D, Liu B, Wang J. Measurement and gender-specific analysis of user publishing characteristics on MySpace. IEEE Netw. 2010;24:38–43.
25. Asur S, Huberman BA. Predicting the Future With Social Media. WI-IAT '10 Proc 2010 IEEE/WIC/ACM Int Conf Web Intell Intell Agent Technol. 2010;429–99.
26. Motoyama M, Voelker GM, Savage S. Measuring Online Service Availability Using Twitter. WOSN'10 Proc 3rd Conf Online Soc networks. 2010;13.
27. Mislove A, Lehmann S, Ahn Y-Y, Onnela J-P, Rosenquist JN. pulse of the nation us mood throughout the day inferred from twitter. 2013;
28. Heaivilin N, Gerbert B, Page J, Gibbs J. Public health surveillance of dental pain via Twitter. J Dent Res. 2011;90:1047–51.
29. Bosley JC, Zhao NW, Hill S, Shofer FS, Asch DA, Becker LB, et al. Decoding twitter: Surveillance and trends for cardiac arrest and resuscitation communication. Resuscitation. 2013;84:206–12.

30. Paul MJ, Dredze M. You Are What You Tweet: Analyzing Twitter for Public Health. Fifth Int AAAI Conf Weblogs Soc Media. 2011;265–72.
31. Gomide J, Veloso A, Meira W, Almeida V, Benevenuto F, Ferraz F, et al. Dengue surveillance based on a computational model of spatio-temporal locality of Twitter. Proc 3rd Int Web Sci Conf - WebSci '11. New York, New York, USA, New York, USA: ACM Press; 2011. p. 1–8. Available from: http://dl.acm.org/citation.cfm?doid=2527031.2527049. Accessed 3 March 2019.
32. Signorini A, Segre AM, Polgreen PM. The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic. PLoS One. 2011.
33. Chew C, Eysenbach G. Pandemics in the age of Twitter: Content analysis of tweets during the 2009 H1N1 outbreak. PLoS One. 2010;5:361–7.
34. Lampos V, Cristianini N. Tracking the flu pandemic by monitoring the social web. 2nd Int Work Cogn Inf Process. Ieee; 2010;411–6.
35. Aramaki E. Twitter Catches The Flu: Detecting Influenza Epidemics using Twitter. Proc 2011 Conf Empir Methods Nat Lang Process. 2011:1568–76.
36. Achrekar H. Social Network Enabled Flu Trends.
37. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature. 2009;457:1012–4.
38. Flu Trackers. 2014. Available from: https://flutrackers.com/forum. Accessed 3 March 2019.
39. Flusurvey. 2014; Available from: https://flusurvey.org.uk. Accessed 3 March 2019.
40. Polgreen PM, Chen Y, Pennock DM, Nelson FD. Using internet searches for influenza surveillance. Clin Infect Dis. 2008;47:1443–8.
41. Hulth A, Rydevik G, Linde A. Web queries as a source for syndromic surveillance. PLoS One. 2009;4:e4378.
42. Paul MJ, Dredze M, Broniatowski D. Twitter Improves Influenza Forecasting. PLoS Curr. 2014; Available from: http://currents.plos.org/outbreaks/?p=39911. Accessed 3 March 2019.
43. Broniatowski DA, Paul MJ, Dredze M. National and Local Influenza Surveillance through Twitter: An Analysis of the 2012–2013 Influenza Epidemic. Preis T, editor. PLoS One. 2013;8:e83672. Available from: http://dx.plos.org/10.1371/journal.pone.0083672
44. Balakrishnan V. System identification: theory for the user (second edition). Automatica. 2002;38:375–8.
45. Ramesh K, Aziz N, Shukor A. R. S. Development of NARX Model for Distillation Column and Studies on Effect of Regressors. J Appl Sci. 2008;8:1214–20.
46. Cajueiro E, Kalid R, Schnitman L. Using NARX model with wavelet network to inferring the polished rod position. Int J Math Comput Simul. 2012;6.
47. Zhang QJ, Gupta KC, Devabhaktuni VK. Artificial neural networks for RF and microwave design - From theory to practice. IEEE Transactions on Microwave Theory and Techniques. 2003;51(4):1339–50. https://doi.org/10.1109/TMTT.2003.809179.
48. Sandoval L. Structure of a global network of financial companies based on transfer entropy. Entropy. 2014;16:4443–82. Available from: http://www.mdpi.com/1099-4300/16/8/4443/. Accessed 3 March 2019.
49. Kendall M. Rank correlation methods. London Griffin. 1970.